SAND2011-0520P

Algebraic Techniques for Multilingual Document Clustering

Brett W. Bader, Ph.D.
Sandia National Laboratories
http://www.sandia.gov/~bwbader

January 25, 2011

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Acknowledgements

- Peter Chew* (computational linguistics/data)
- Ron Oldfield (HPC/architecture)
- Philip Kegelmeyer (machine learning)
- Sue Medeiros
- Alla Fishman
- Tim Shead
- Nathan Fabien
- Tamara Kolda
- Jon Stearley
- George Davidson
- Craig Ulmer
- Todd Kordenbrock
- Stephen Verzi

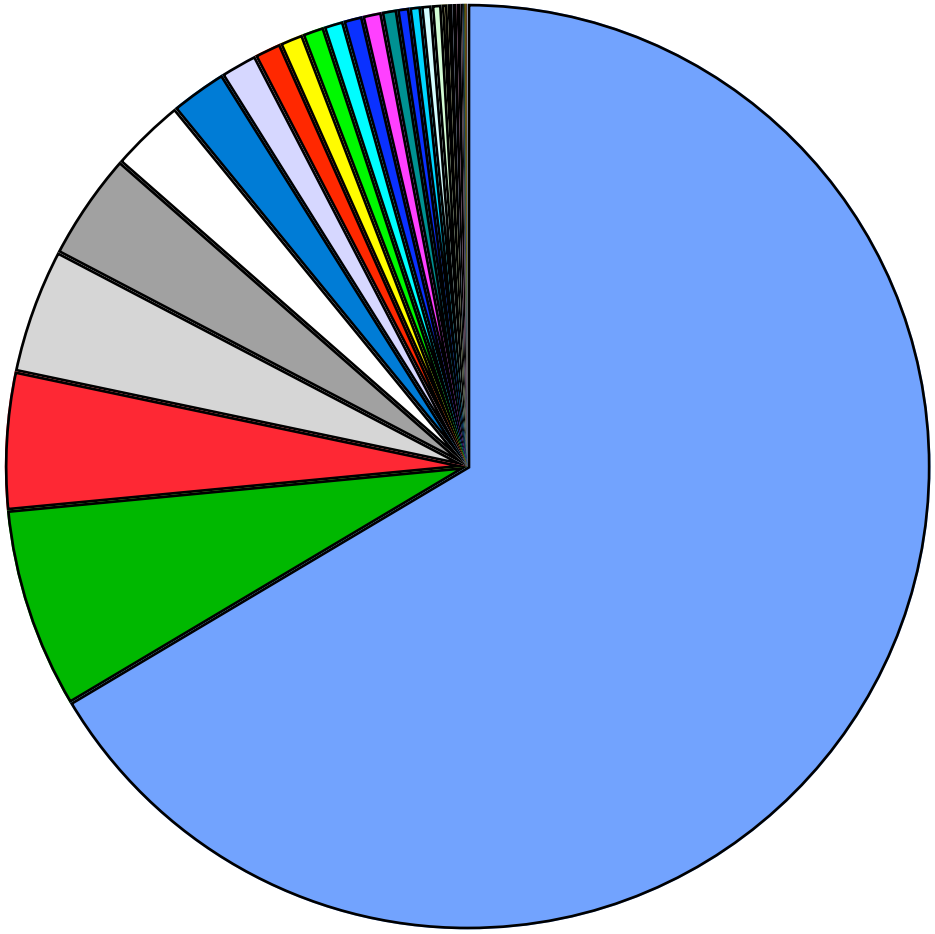
- Ahmed Abdelali
- Stephen Helmreich

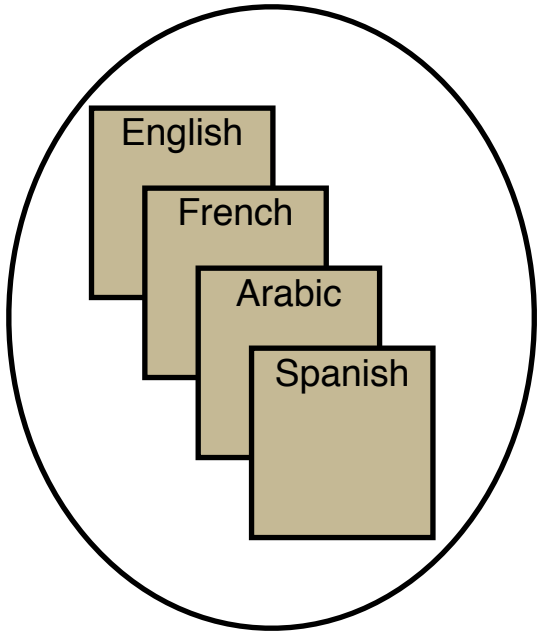
* current affiliation: Galisteo Consulting

SNL has developed multilingual techniques to analyze documents across multiple languages

- "Translate" new documents into a language-independent concept space, which is useful for:
  - Document clustering
  - Translation triage (i.e., translate documents in clusters of interest)
  - Ideological classification (e.g., hostile to democracy)
  - Multilingual sentiment analysis

Languages on the web (ca. 2006)





Sandia National Laboratories

Bag of Words/Vector Space Model

example from (Berry, Drmac, Jessup, 1999)

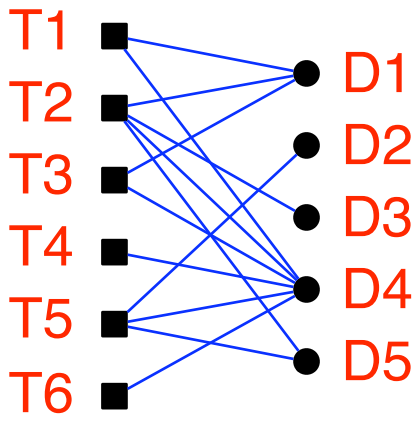
Documents

D1: How to Bake Bread Without Recipes
D2: The Classic Art of Viennese Pastry
D3: Numerical Recipes: The Art of Scientific Computing
D4: Breads, Pastries, Pies and Cakes: Quantity Baking Recipes
D5: Pastry: A Book of Best French Recipes

Terms

T1: bak(e,ing)
T2: recipes
T3: bread
T4: cake
T5: pastr(y,ies)
T6: pie

Bipartite graph



Term-by-doc (adjacency) matrix

$$\hat{A} = \begin{array}{c} \\ \text{T1} \\ \text{T2} \\ \text{T3} \\ \text{T4} \\ \text{T5} \\ \text{T6} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{D1} & \text{D2} & \text{D3} & \text{D4} & \text{D5} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Key concepts

- Bag of words
- Stop words
- Stemming
- Vector space model
- Scaling for information content

Design Goals

- Allow as many languages as possible

- Rely solely on statistical analysis of a corpus, no language experts
  - No stemming
  - No stoplists, keep all terms

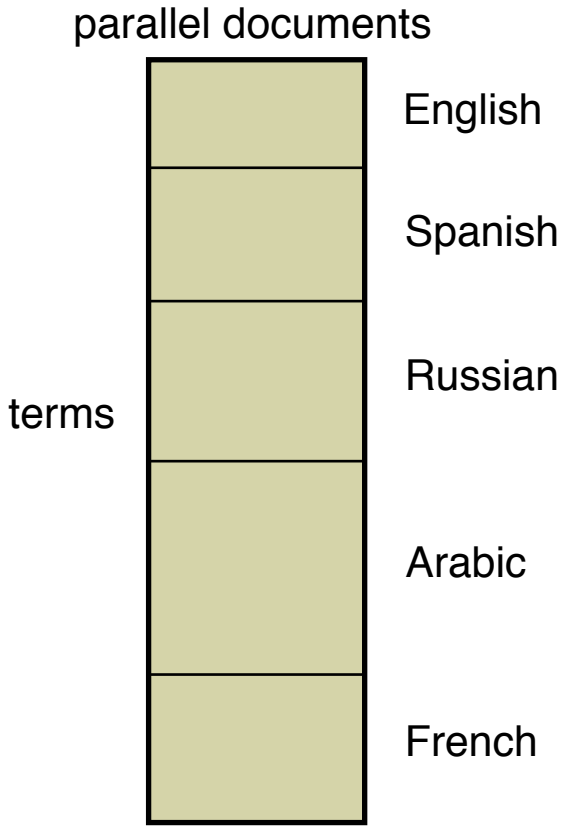
require human labor/expertise

Language expertise in our techniques,
but no language expertise required to use

Term-Document Matrix

Term-by-doc matrix for all languages

Rosetta Stone



Look for co-occurrence of terms in the same documents and across languages to capture latent concepts

- Approach is not new: pairs of languages in Latent Semantic Analysis (LSA)
  - English and French (Landauer & Littman, 1990)
  - English and Greek (Young, 1994)
- *Multi-parallel* corpus is new

Bible as a 'Rosetta Stone'

- The Bible has been translated carefully and widely
  - 451 complete & 2479 partial translations
- Verse aligned

Sandia's database: 54 languages: >99% coverage of web

Afrikaans	Estonian	Norwegian
Albanian	Finnish	Persian (Farsi)
Amharic	French	Polish
Arabic	German	Portuguese
Aramaic	Greek (New Testament)	Romani
Armenian Eastern	Greek (Modern)	Romanian
Armenian Western	Hebrew (Old Testament)	Russian
Basque	Hebrew (Modern)	Scots Gaelic
Breton	Hungarian	Spanish
Chamorro	Indonesian	Swahili
Chinese (Simplified)	Italian	Swedish
Chinese (Traditional)	Japanese	Tagalog
Croatian	Korean	Thai
Czech	Latin	Turkish
Danish	Latvian	Ukrainian
Dutch	Lithuanian	Vietnamese
English	Manx Gaelic	Wolof
Esperanto	Maori	Xhosa

Bible as Parallel Corpus

5 languages for training and testing

Translation	Terms	Total Words
English (King James)	12,335	789,744
French (Darby)	20,428	812,947
Spanish (Reina Valera 1909)	28,456	704,004
Russian (Synodal 1876)	47,226	560,524
Arabic (Smith Van Dyke)	55,300	440,435

- Languages convey information in different number of words

Isolating language ⟷ Synthetic language

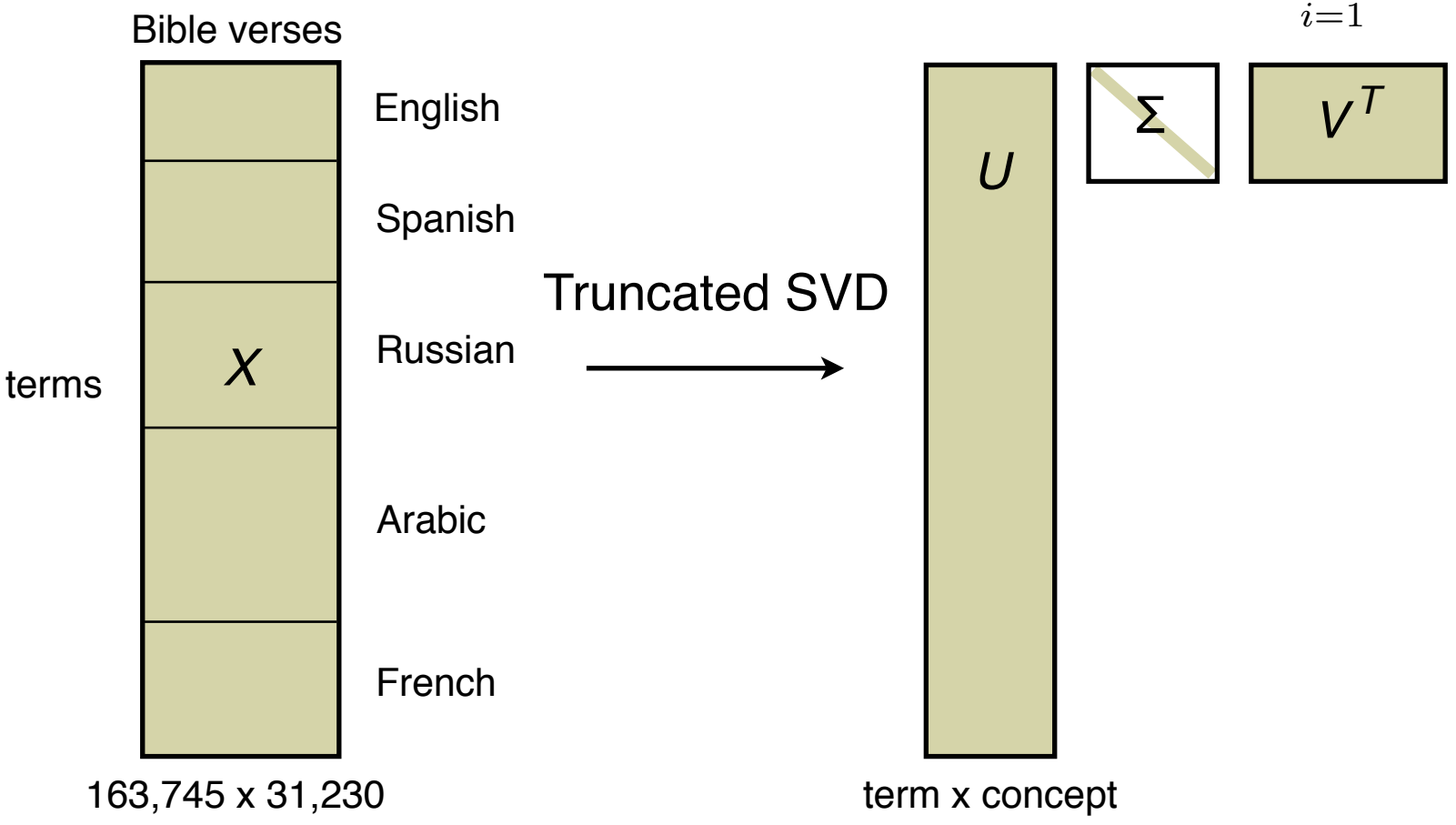
Example of Statistical Differences

	Text	word count	% of total
AR	في البدء خلق الله السموات والارض.	6	14
EN	In the beginning God created the heavens and the earth.	10	24
FR	Au commencement Dieu créa les cieux et la terre.	9	21
RU	В начале сотворил Бог небо и землю.	7	17
ES	En el principio crió Dios los cielos y la tierra.	10	24
TOTAL		**42**	**100**

Multilingual Latent Semantic Analysis

Term-by-verse matrix
for all languages

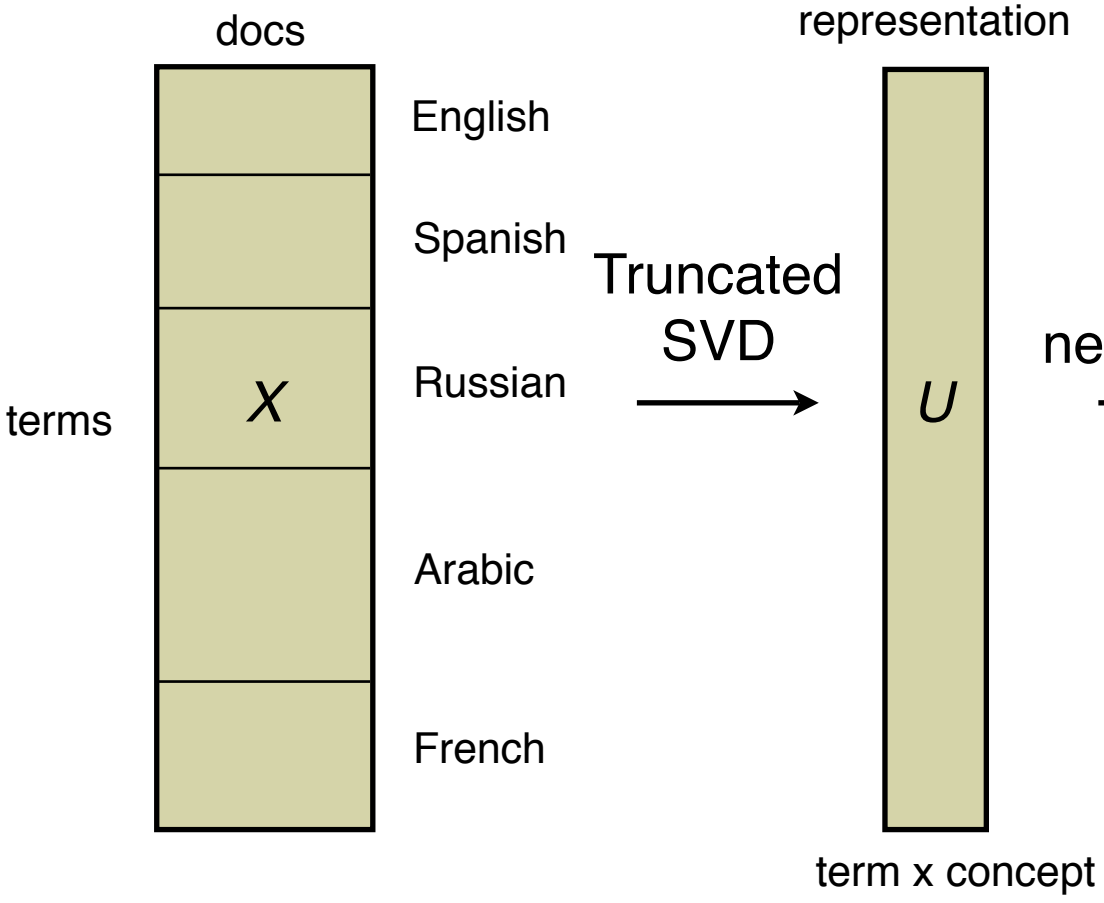
$$X_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$



SVD allows both terms and documents to be
mapped to a single set of cross-language concepts

Multilingual Latent Semantic Analysis

Term-by-doc matrix for all languages



"Translate" new documents into a small number of language-independent features

dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

Document feature vector

Applications

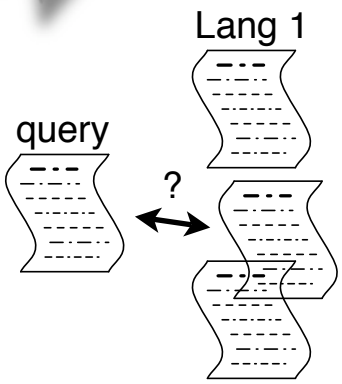
- cross-language retrieval
- pairwise similarities for clustering
- machine learning applications

Verification and Validation

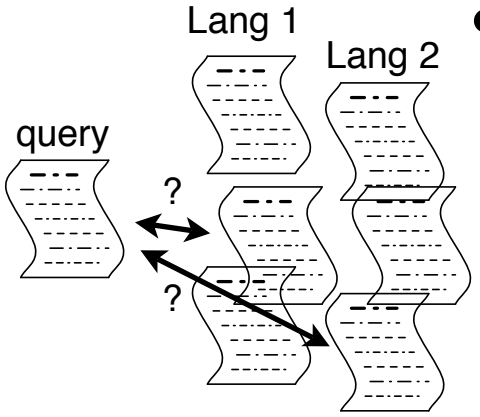
- Bible as training set
- Quran as test set

- Quran is translated into many languages, just like the Bible
  - Multi-parallel corpus
  - Ground truth
  - 114 suras (or chapters)
  - More variation across translations => harder IR task

Performance Metrics



- Average precision at 1 document (P1)
  - Equals the percentage of times the translation of the query ranked highest
  - Essentially, P1 measures success in retrieving documents when the source and target languages are specified



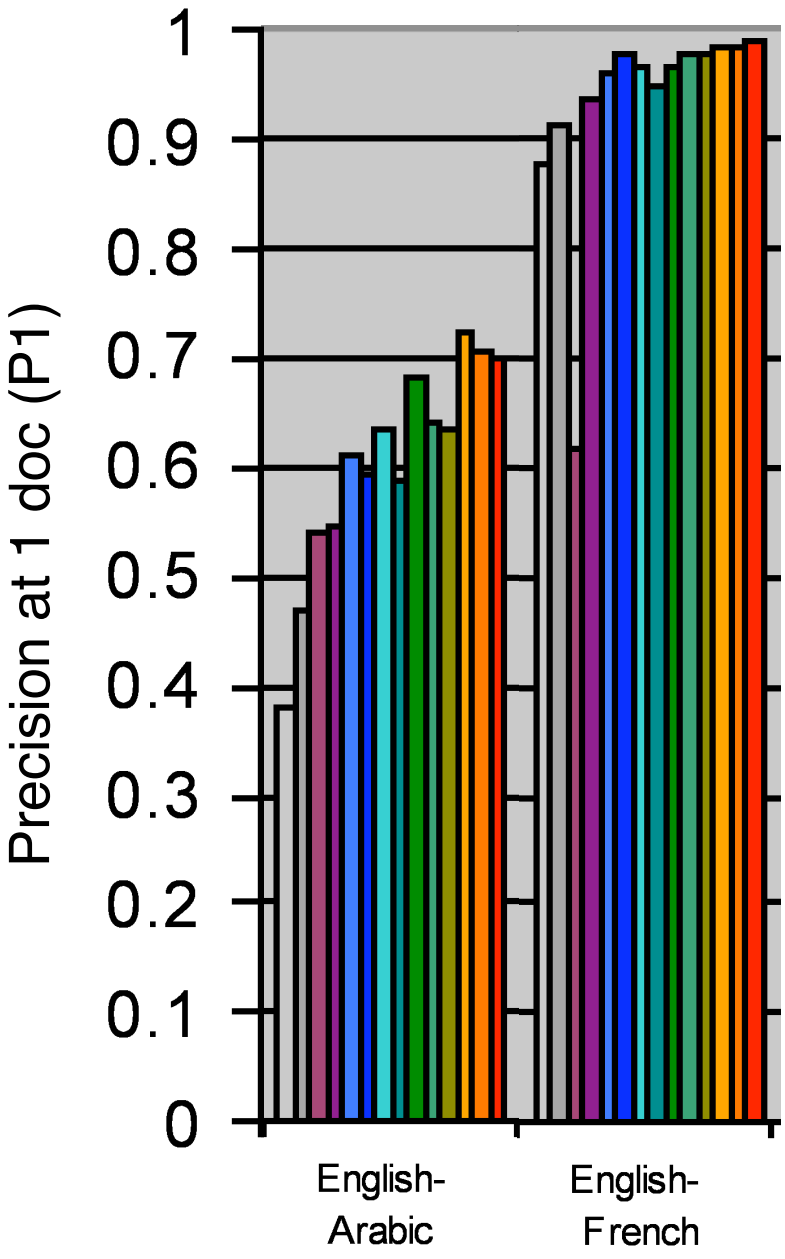
- Average multilingual precision at 5 (or n) documents (MP5)
  - The average percentage of the top 5 documents that are translations of the query document
  - Calculated as an average for all queries & all languages
  - Essentially, MP5 measures success in multilingual clustering

- Standard measures from information retrieval but adapted for multiple languages
- Striving for 90% MP5

Multilingual LSA

(Chew and Abdelali, 2007)

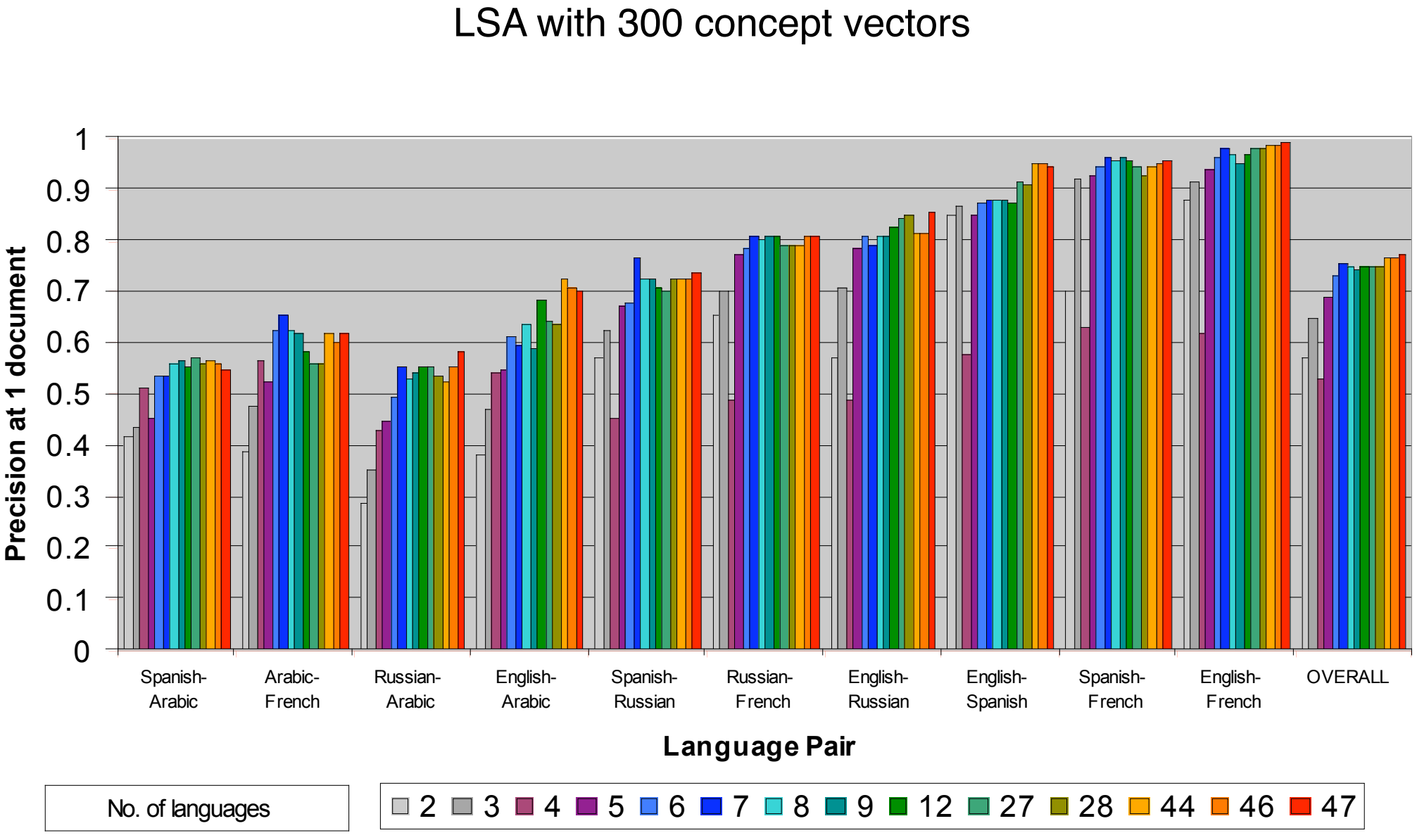
LSA with 300 concept vectors



- More training languages = better results
  - Train on 2 to 47 languages
- Some languages are harder than others
  - e.g., French vs. Arabic

More languages = Better results

(Chew and Abdelali, 2007)



Sandia National Laboratories

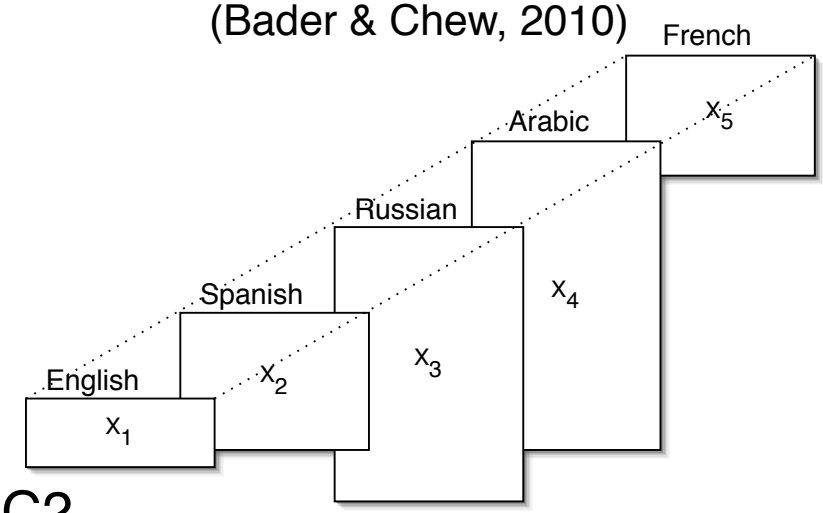
Improved CLIR Methods & Results

Overall Results

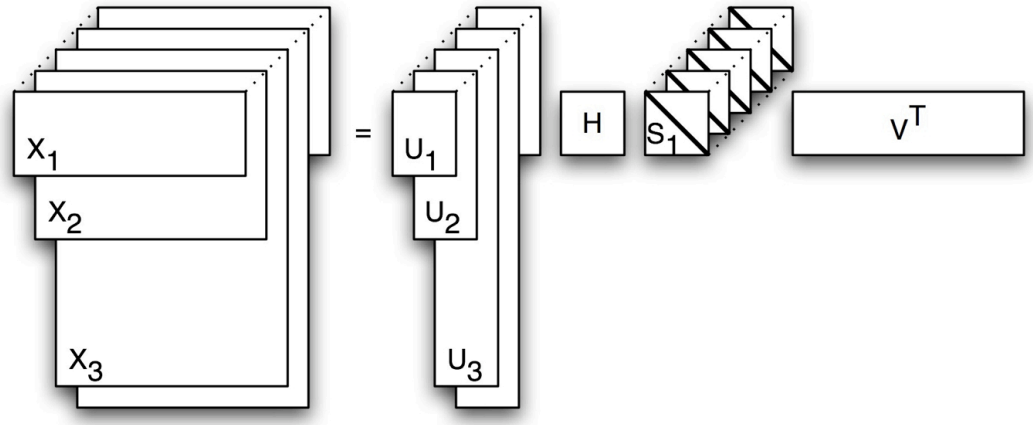
Method	MP5
SVD/LSA (α=1)	26.1%
SVD/LSA (α=1.8)	65.5%
Tucker1	71.3%
PARAFAC2	78.5%
LSATA	80.7%

Early on, documents tended to cluster more by language than by topic

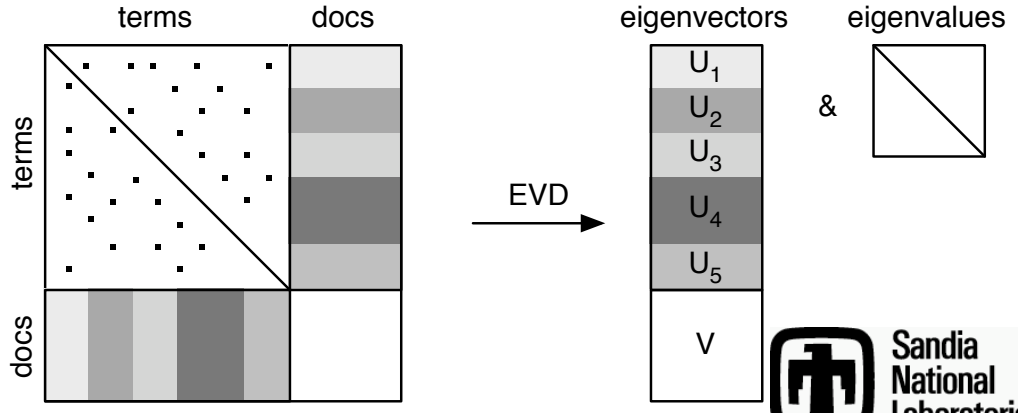
(Bader & Chew, 2010)



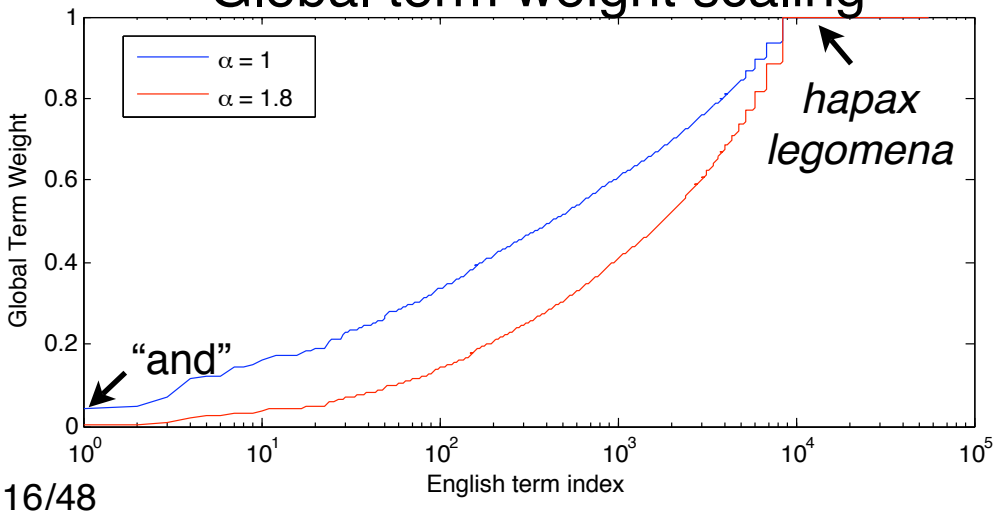
PARAFAC2



LSATA



Global term weight scaling

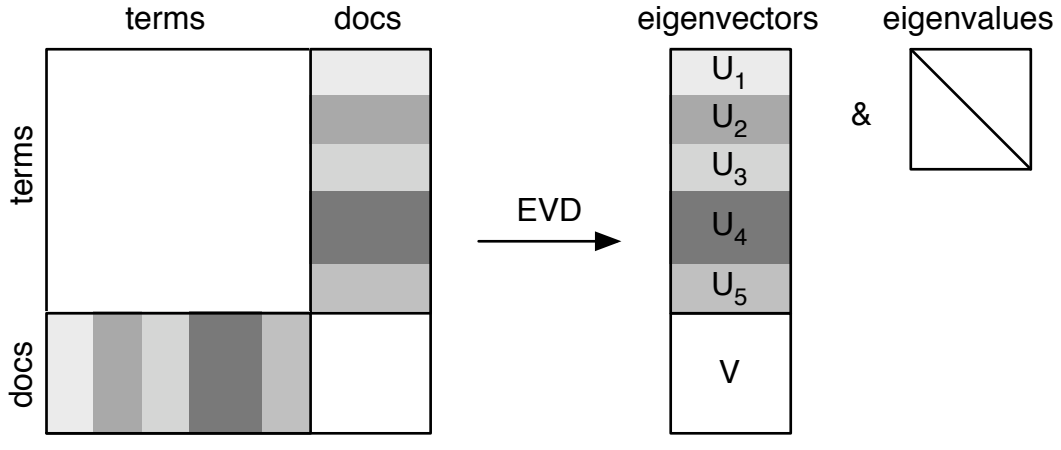


Sandia National Laboratories

Calculating the SVD in LSA

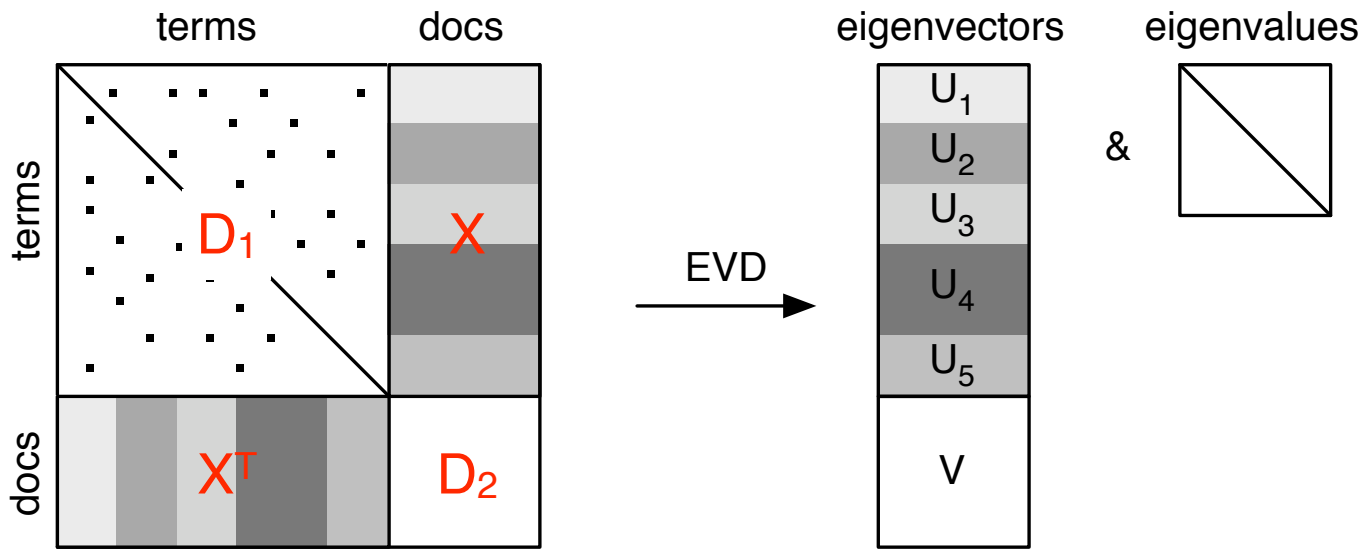
SVD: $X = U\Sigma V^T$

Matrix		eigenvectors		eigenvalues
XX^T	$\longrightarrow$	U	&	Σ^2
X^TX	$\longrightarrow$	V	&	Σ^2
$\begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix}$	$\longrightarrow$	$\frac{1}{\sqrt{2}}\begin{pmatrix} U_+ & \sqrt{2}U_0 & -U_+ \\ V & 0 & V \end{pmatrix}$	&	$\begin{pmatrix} \Sigma & & \\ & 0 & \\ & & -\Sigma \end{pmatrix}$



LSA with Term Alignments (LSATA)

(Bader and Chew, 2008)



Add term-alignment information into the diagonal block to strengthen the co-occurrence information that LSA normally finds in the parallel corpus via the SVD.

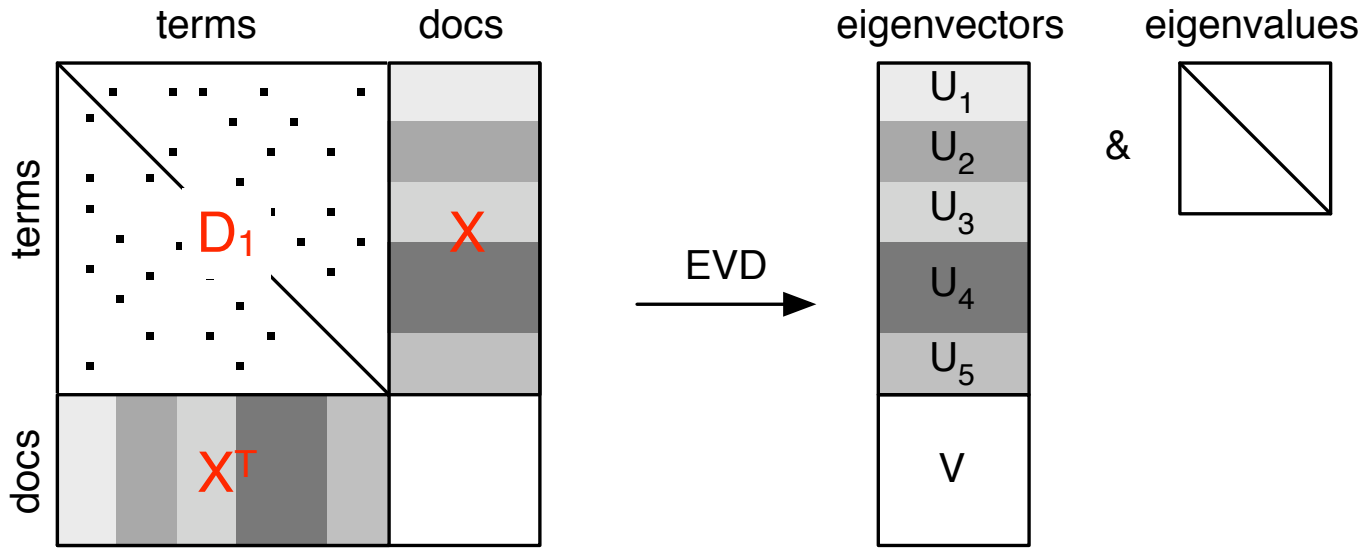
Possibilities for D_1:

- Binary entries

  D_{ij} = 1 if the pair (i,j) occurs in a dictionary, 0 otherwise

- Pairwise mutual information (as in statistical machine translation - SMT)

Algorithmic Interpretation



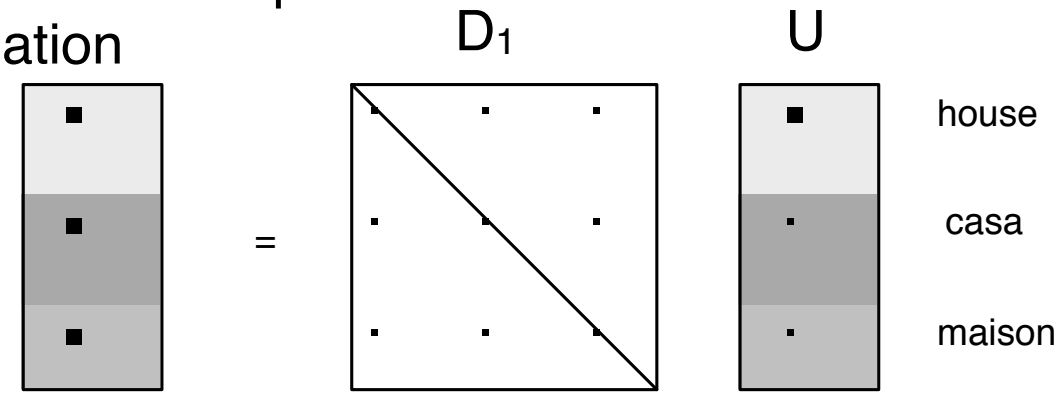
Power method:

$$U_{new} = D_1 U + XV$$

$$V_{new} = X^T U$$

Reinforce term-term relationships from external information

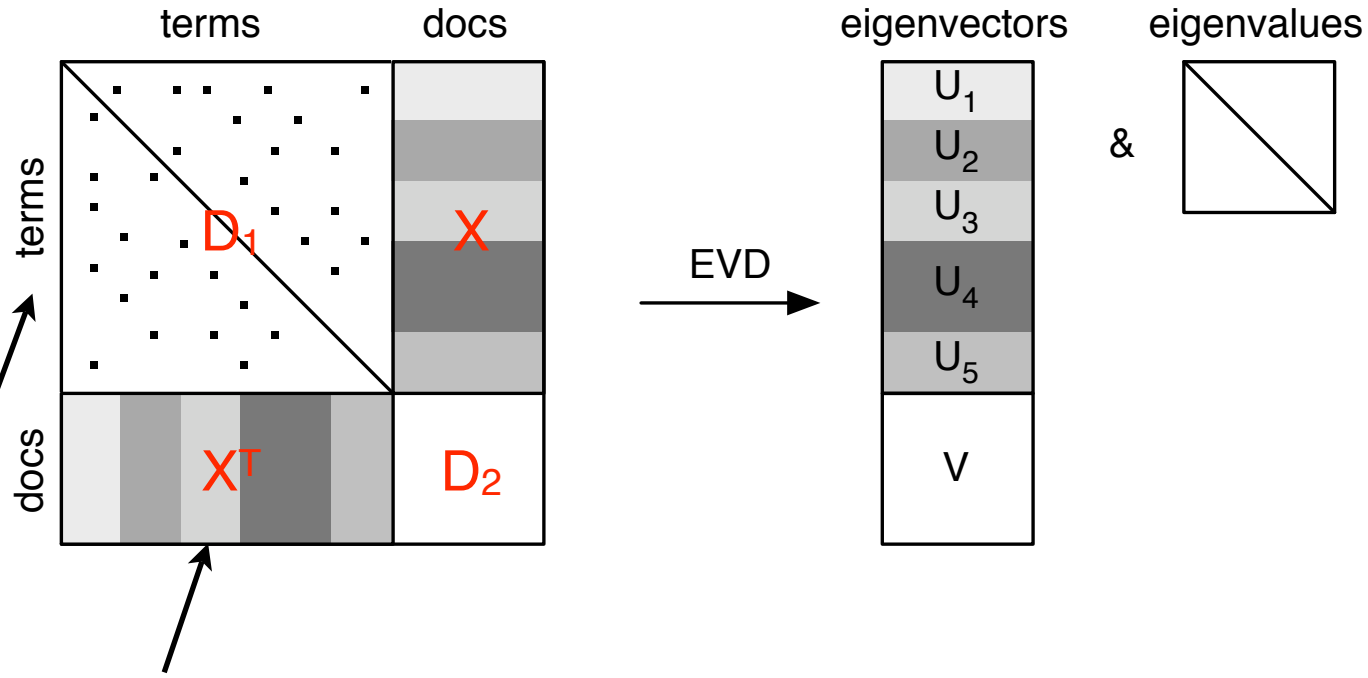
Standard relationship in LSA



Relationship between *house, casa,* and *maison* is strengthened

Matrix Scaling with CL Roots

LMSA with Term-Alignments



Pointwise Mutual Information (PMI)

$$X_{i,j} = \log\left(\frac{p(i,j)}{p(i)\cdot p(j)}\right)$$

Mutual Information

$$I(A,B) = \sum_{i\in A}\sum_{j\in B} p(i,j)\log\left(\frac{p(i,j)}{p(i)\cdot p(j)}\right)$$

Multilingual precision at 5 documents: 80.7%

Language Morphology

Translation	Terms	Total Words
English (King James)	12,335	789,744
Arabic (Smith Van Dyke)	55,300	440,435

Languages convey information in different number of words

Morphemes are ‘the smallest individually meaningful elements in the utterances of a language’
(Hockett, 1958)

Language Morphology

Translation	Terms	Total Words
English (King James)	12,335	789,744
Arabic (Smith Van Dyke)	55,300	440,435

Languages convey information in different number of words

Isolating language ⟷ Synthetic language

Chinese — Quechua, Inuit (Eskimo)

- Isolating language: One morpheme per word
  - e.g., "He travelled by hovercraft on the sea." Largely isolating, but travelled and hovercraft each have two morphemes per word. (Wikipedia)

Language Morphology

Translation	Terms	Total Words
English (King James)	12,335	789,744
Arabic (Smith Van Dyke)	55,300	440,435

Languages convey information in different number of words

Isolating language ⟷ Synthetic language

Chinese — Quechua, Inuit (Eskimo)

- Isolating language: One morpheme per word
  - e.g., "He travelled by hovercraft on the sea." Largely isolating, but travelled and hovercraft each have two morphemes per word. (Wikipedia)
- Synthetic language: High morpheme-per-word ratio
  - German: *Aufsichtsratsmitgliederversammlung* => "On-view-council-with-limbs-gathering" meaning "meeting of members of the supervisory board". (Wikipedia)
  - Chulym: *Aalychtypiskem* => "I went out moose hunting"
  - Yup'ik Eskimo: *tuntussuqatarniksaitengqiggtuq* => "He had not yet said again that he was going to hunt reindeer." (Payne, 1997)

Morphological Tokenization

Our hypothesis: if the terms were morphemes, not words or stems, the results of IR would be improved.

- Two approaches:
  - Tokenization based on mutual information of character n-grams
  - Unsupervised learning of morphology from a corpus based on Minimum Description Length (Goldsmith, 2001)
    - Linguistica (open source)
- Generalizable to new languages
- Unsupervised

Tokenization from n-gram mutual information

(Chew, Bader, Abdelali, 2008)

- Consider all possible tokenizations
  - "walked" --> walked, w+alked, wa+lked, ..., walk+ed, walke+d, ..., w+a+l+k+e+d

- Calculate pointwise mutual information (PMI) of each n-gram individually from the corpus

$$PMI(\text{“}walk\text{”}) = \log\left(\frac{Pr(walk)}{Pr(w)\cdot Pr(a)\cdot Pr(l)\cdot Pr(k)}\right)$$

- Sum the PMI for each tokenization and select the result that is closest to 0

$$Score(walked) = PMI(walked)$$
$$Score(walk + ed) = PMI(walk) + PMI(ed)$$
$$Score(wa + lked) = PMI(wa) + PMI(lked)$$

Sample Tokenization

Wordform	Tokenization
abaissée	*abaissé + e*
abaissées	*abaissé + es*
abaissèrent	*abaiss + èrent*
acceptance	*accept + ance*
acceptation	*accept + ation*
acquaintance	*acquaint + ance*

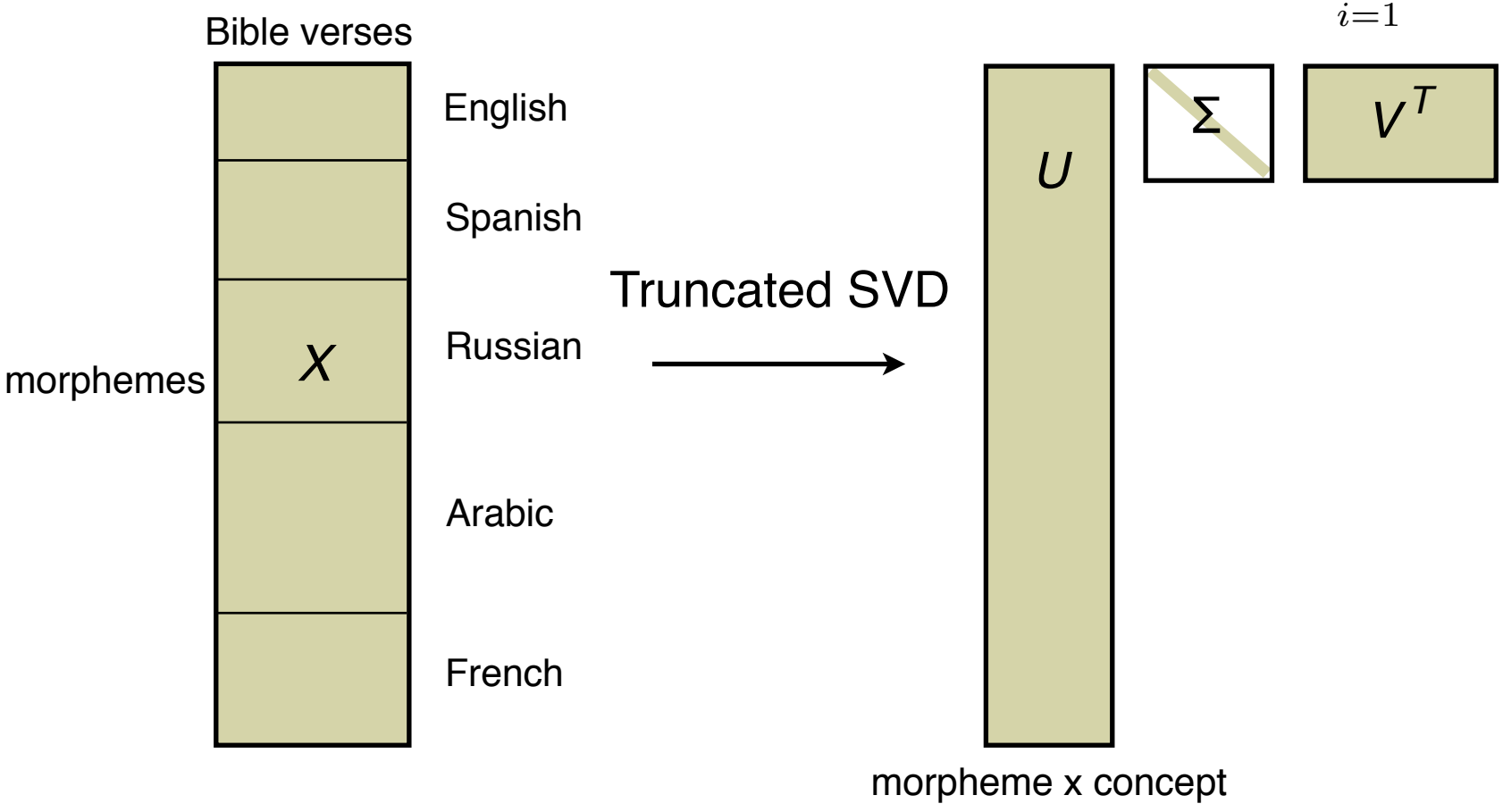
We use these “morphemes”
in place of terms

Latent Morpho-Semantic Analysis (LMSA)

(Chew, Bader, Abdelali, 2008)

Morpheme-by-verse matrix for all languages

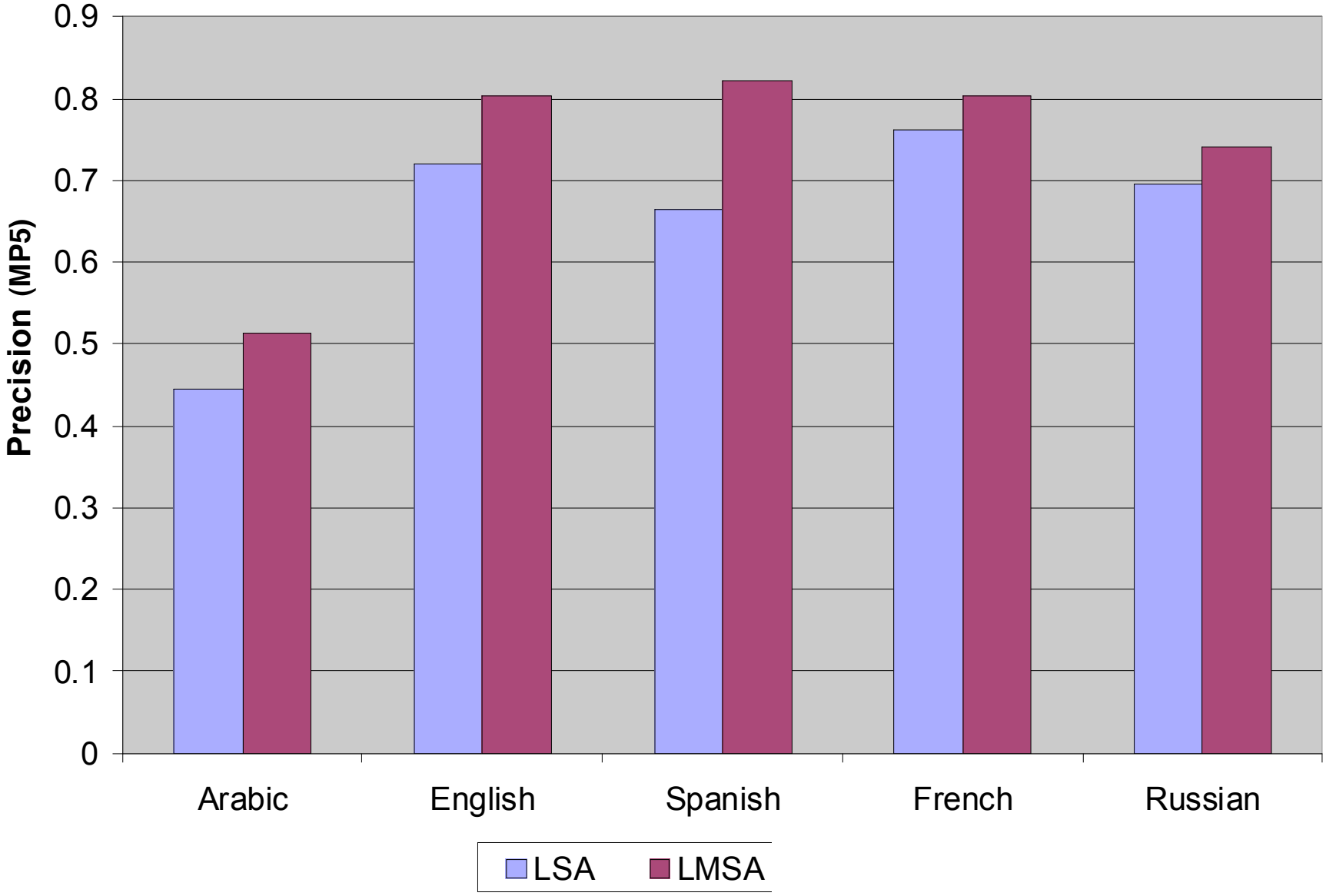
$$X_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$



- Fewer morphemes than terms
- X matrix is smaller but denser

Comparison by Language

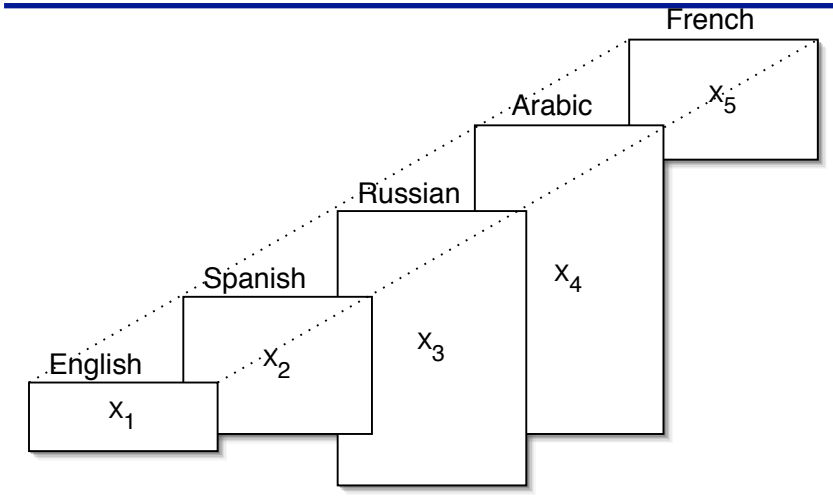
(Chew, Bader, Abdelali, 2008)



Statistically significant improvements
at $p < 0.001$

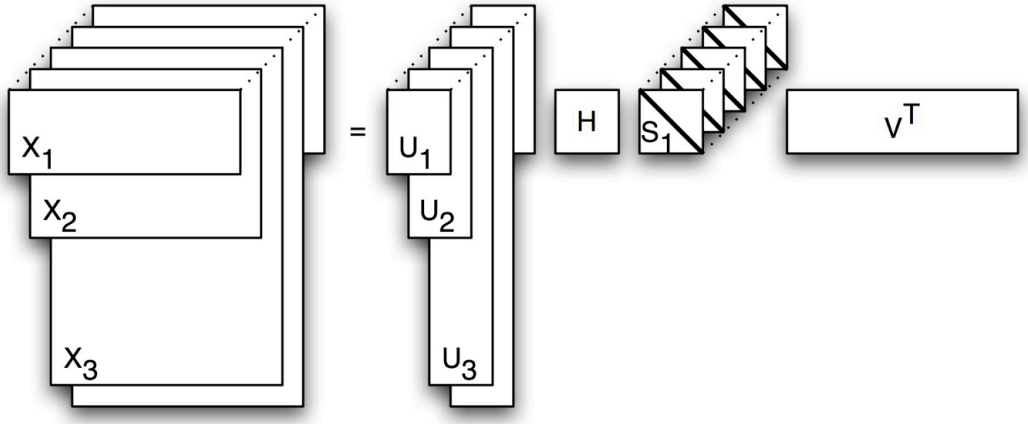
Sandia National Laboratories

Improved CLIR Methods & Results



(Bader & Chew, 2010)

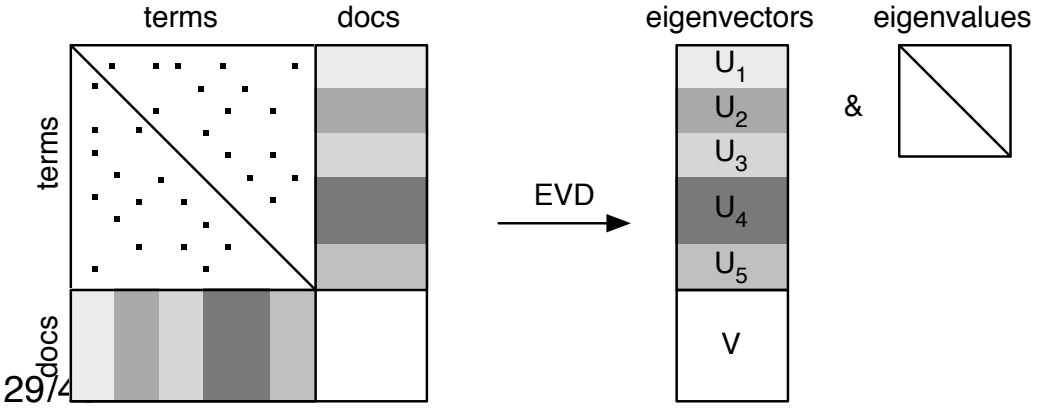
PARAFAC2



Overall Results

Method	MP5
SVD/LSA (α=1)	26.1%
SVD/LSA (α=1.8)	65.5%
Tucker1	71.3%
PARAFAC2	78.5%
LSATA	80.7%
LMSA	73.7%
LMSATA	88.1%

LSATA & LMSATA

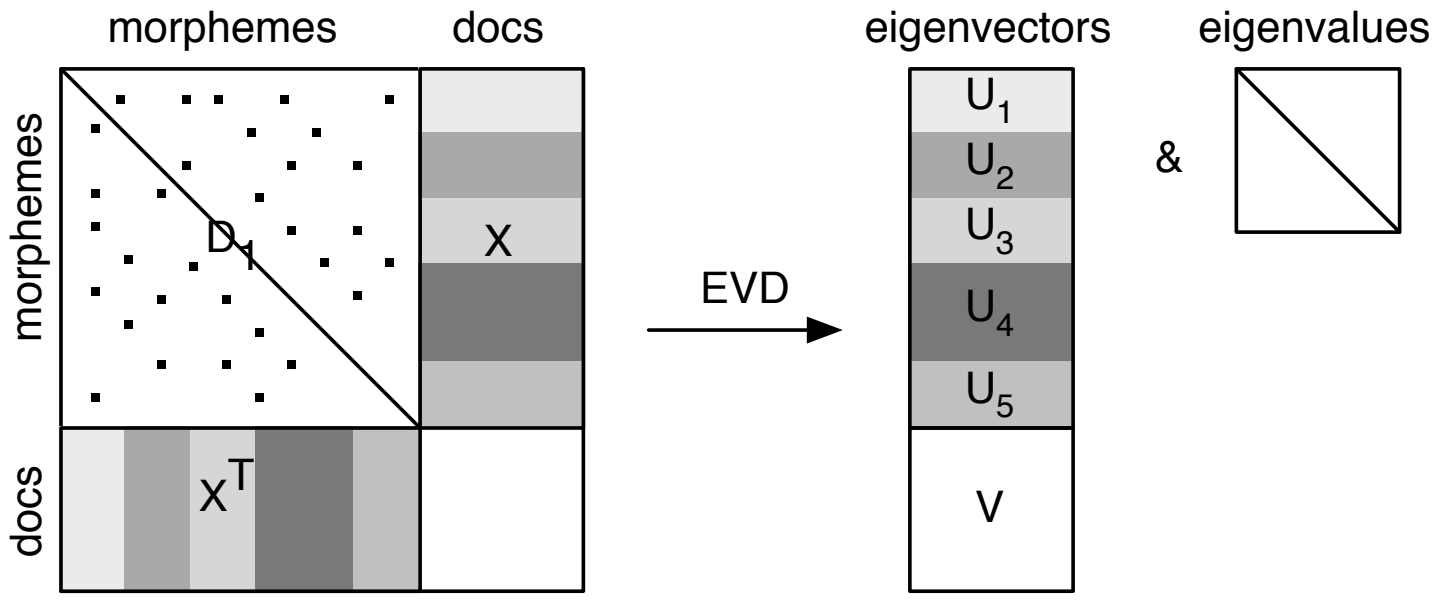


- Early on, documents tend to cluster more by language than by topic
- Morphology represents significant improvement

Sandia National Laboratories

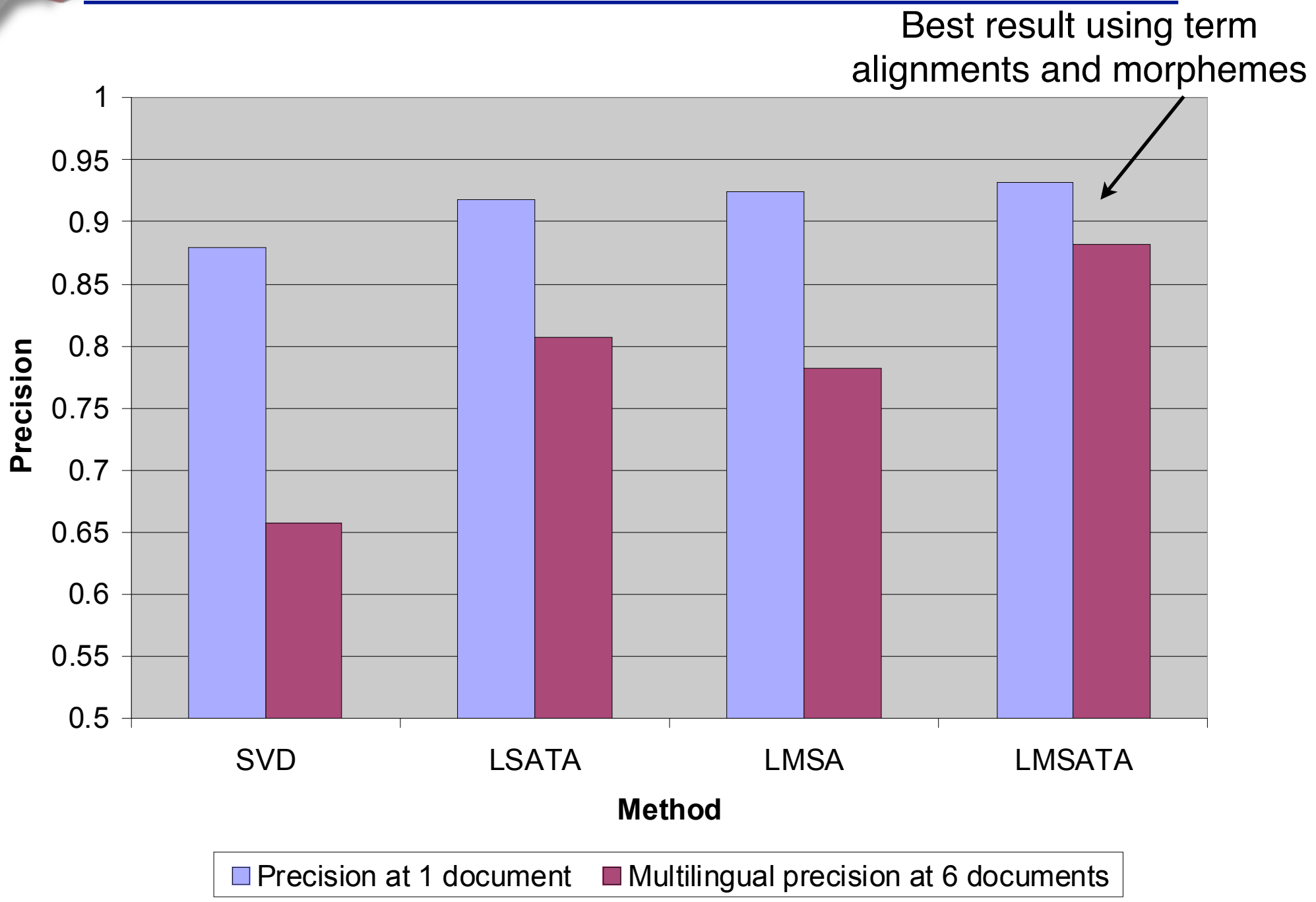
LMSATA: Combine LMSA & LSATA

- Use statistical analysis of character n-grams to get morphemes
- Determine alignment of morphemes for use in LSATA framework



Multilingual precision at 5 documents: 88.1%

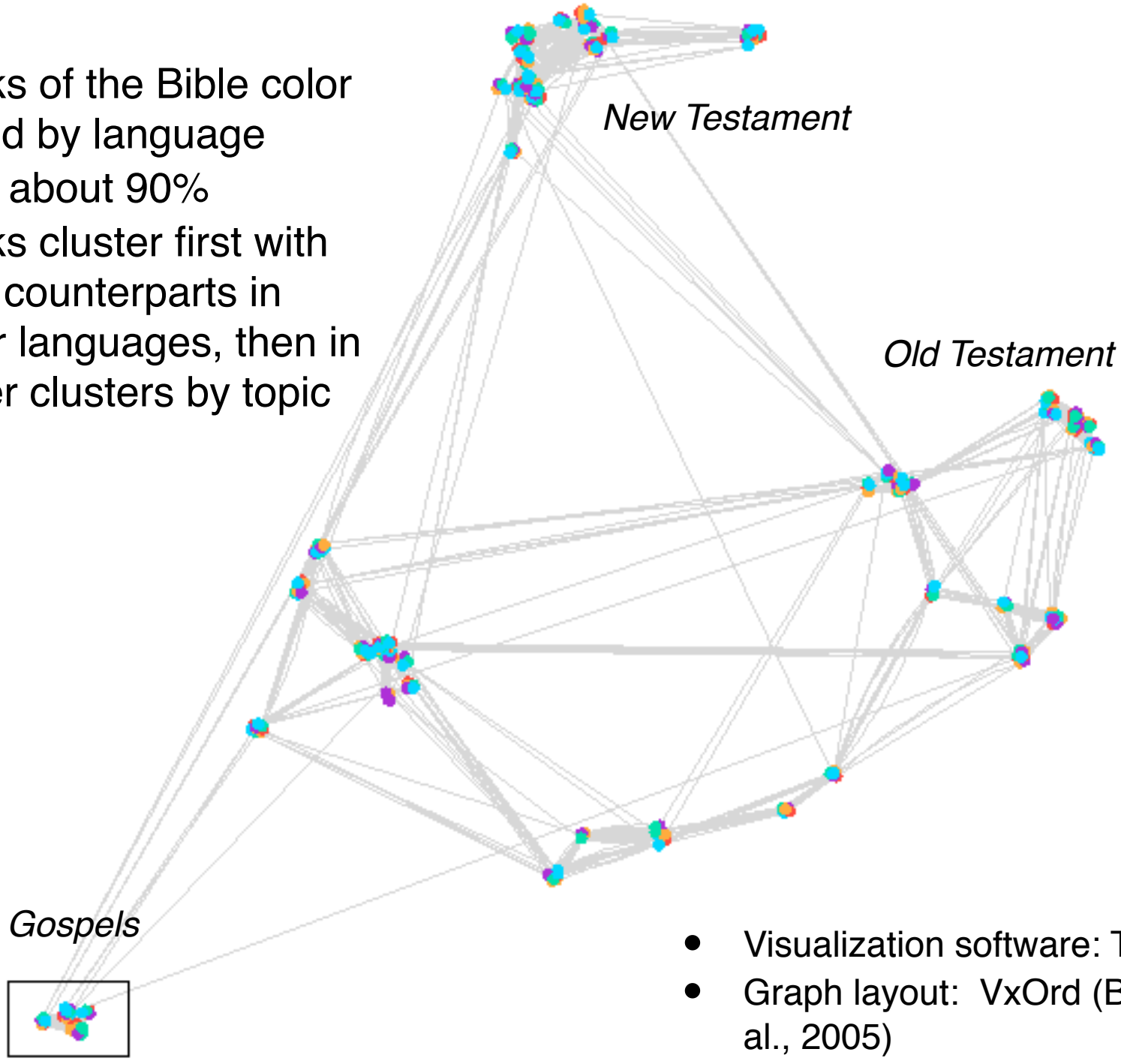
Comparison of LSATA, LMSA, and LMSATA



Sandia National Laboratories

Bible Clustering with LMSATA

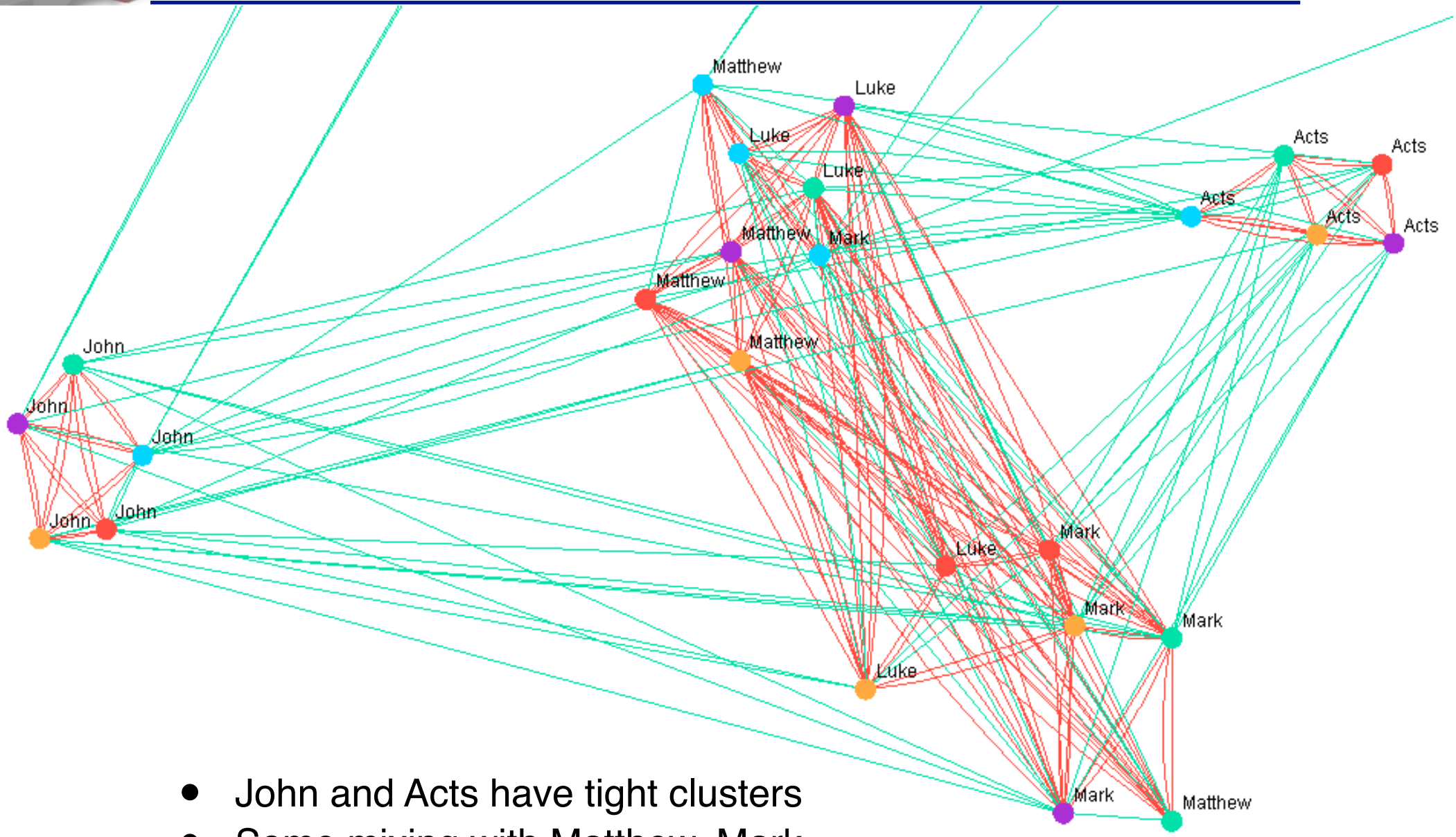
- Books of the Bible color coded by language
- MP5 about 90%
- Books cluster first with their counterparts in other languages, then in larger clusters by topic



- Visualization software: Tamale 1.2
- Graph layout: VxOrd (Boyack et al., 2005)

Sandia National Laboratories

Clustering Close-up



- John and Acts have tight clusters
- Some mixing with Matthew, Mark, Luke (synoptic gospels - share a similar perspective)

Multilingual Clustering is a Great Candidate for HPC

- Scale of Data
  - Millions of elements (Wikipedia, Europarl)
  - Computationally expensive (matrix multiplies for large matrices)
- Time to Solution
  - Interactive control/vis is a motivating factor
  - Focus on “strong scaling” capabilities of HPC platform
- Leveraging Existing Sandia Libraries
  - LMSA for dataset generation
  - Trilinos for computation
  - Titan for visualization
  - Nessie for data services (provides “glue” to integrate systems)

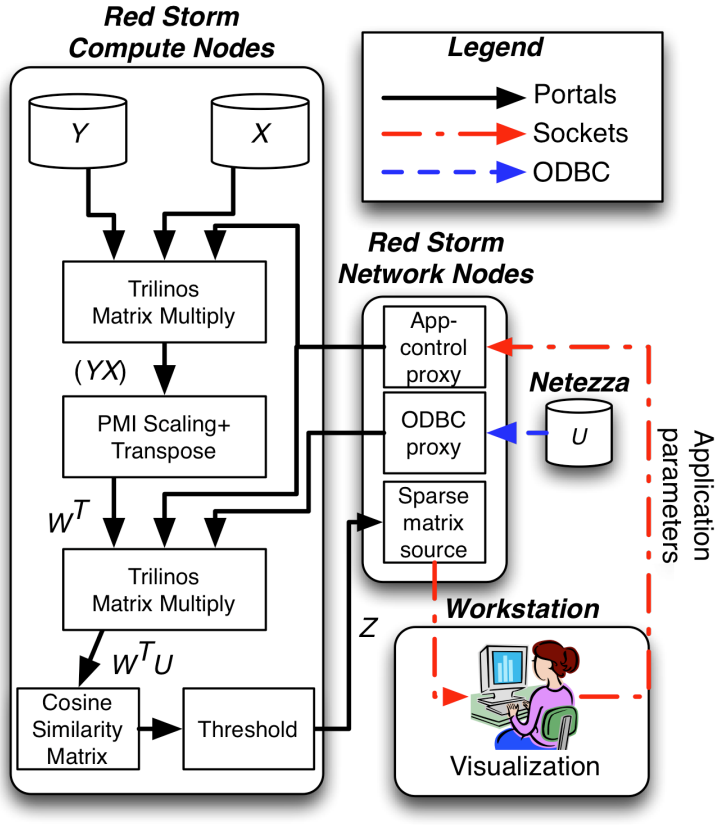
Europarl Corpus

- Extracted from the proceedings of the European Parliament
- Translations in 11 languages
  - French, Italian, Spanish, Portuguese (Romantic)
  - English, Dutch, German, Danish, Swedish (Germanic)
  - Greek
  - Finnish
- Sentence aligned text
- 16 M sentences across 11 languages
- 1,247,832 speeches (including translations)
- 1,249,253 terms (from all 11 languages)

Architectural Challenges

Exploiting specialized architectures

- Red Storm for numerics
- Clusters/Workstations for vis and interactive control
- Data Warehouse Appliances for database functionality



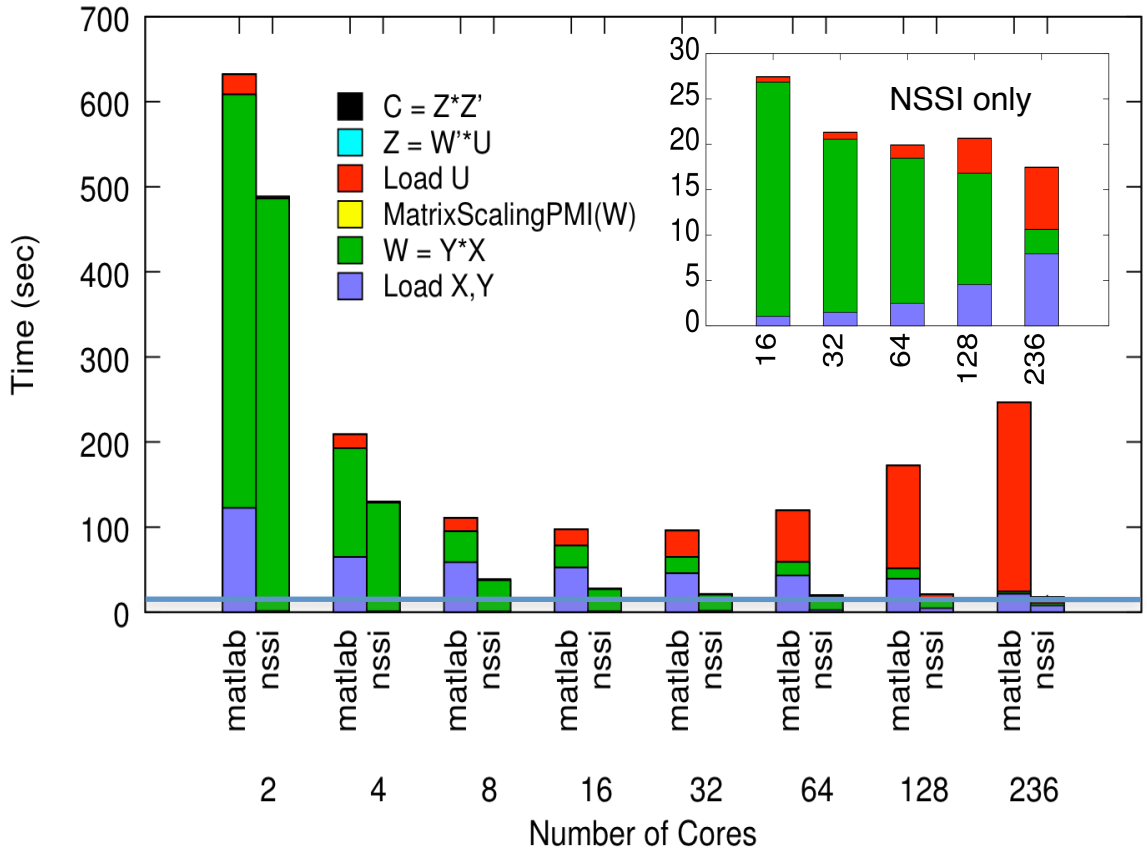
Integrating these systems for interactive jobs has never been done

Sandia National Laboratories

Scaling Challenges for Multilingual Clustering

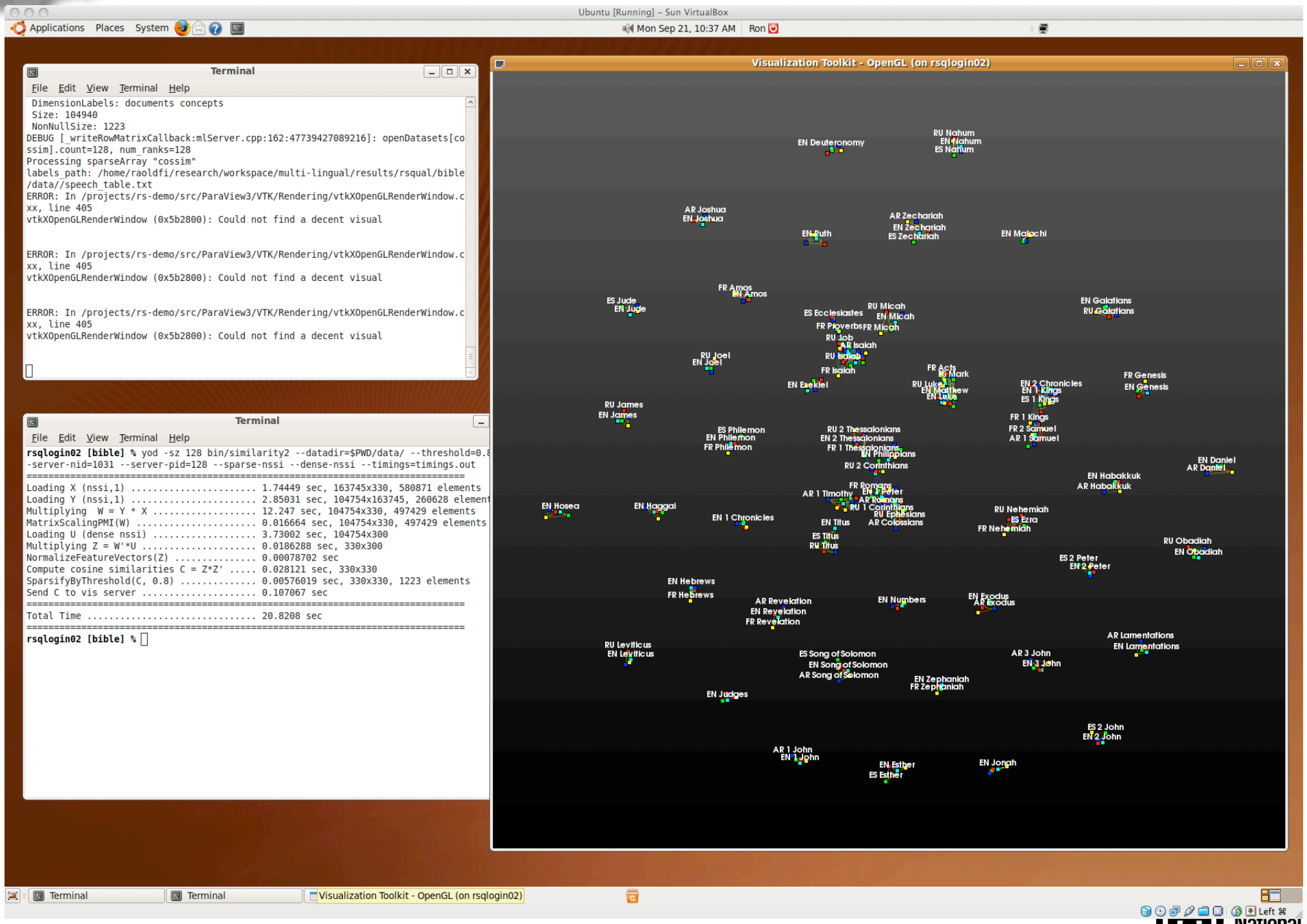
- Strong scaling exposes weaknesses in loading
  - Original methods for loading were not designed for production use.
- Improvements
  - Sparse Reads
    - Keep track of processor mapping information
    - Parallel I/O
  - Dense Reads
    - Convert to binary format
    - Parallel I/O
    - Data ordering
- Status on Red Storm (Cray XT4)
  - 250K docs of Europarl dataset requires 2048 nodes to execute (memory constrained)
  - At 4096 cores, we overwhelm network communication layer when reading input
  - Our target data set has over 1M docs

Performance Results: Bible Dataset



Sandia National Laboratories

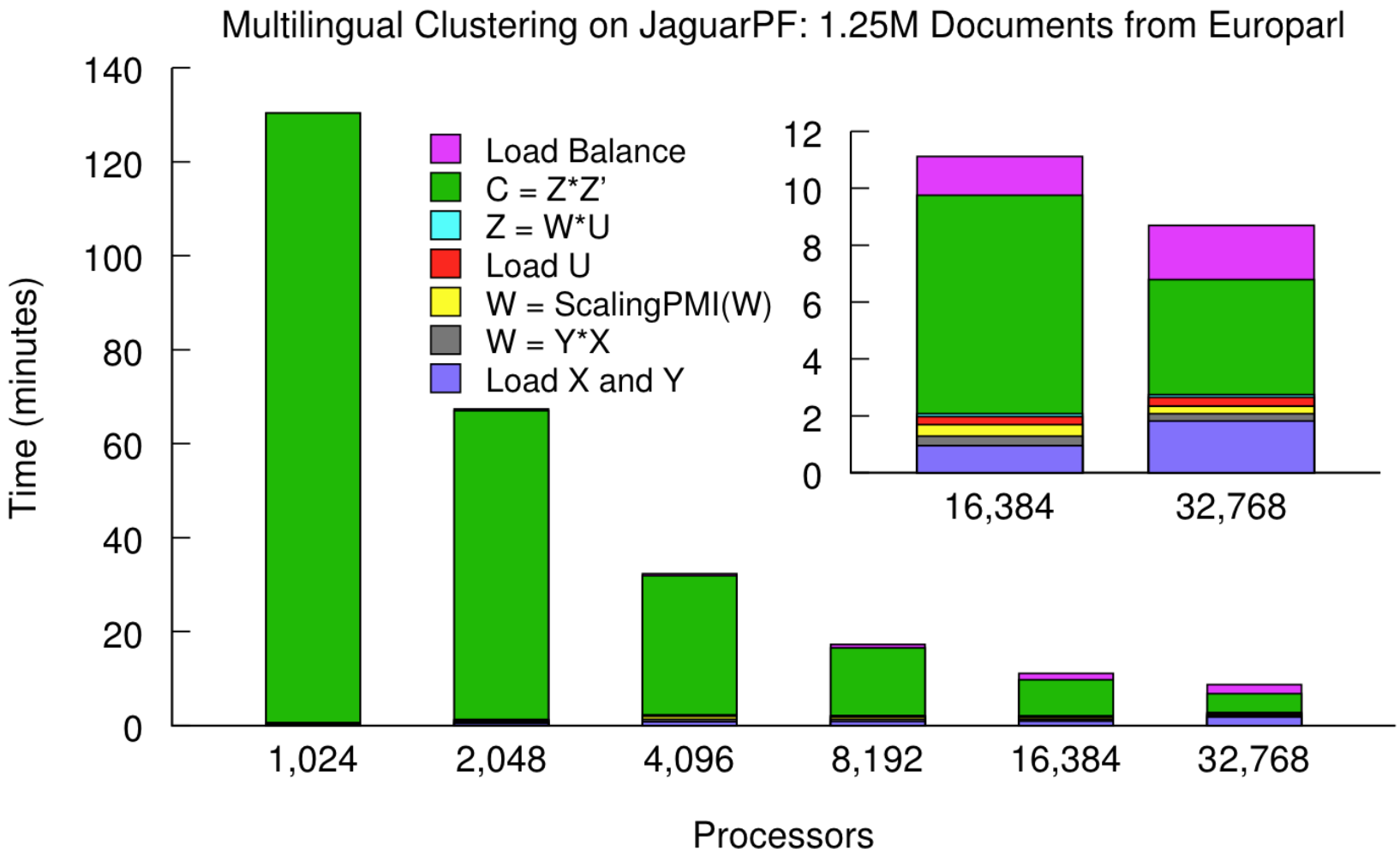
HPC Clustering Demo



National Laboratories

Large-scale Multilingual Clustering

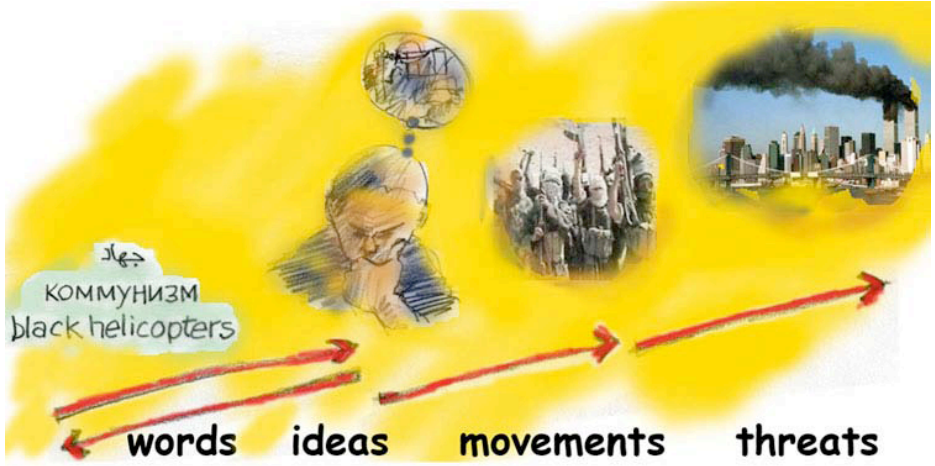
- Performance on JaguarPF (Cray XT5)
  - 1.25M docs of Europarl data set
  - With 32K cores, it takes 470 seconds



Sandia National Laboratories

Predicting Ideology from Document Feature Vectors

(Chew, Kegelmeyer, Bader and Abdelali, 2008)



Hypothesis: there could be a link between *religious* texts and threats

Document feature vector

dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

Assumption: certain sub-regions of the k-dimensional concept space could denote ideological content

Ideological Test Set

(Chew, Kegelmeyer, Bader and Abdelali, 2008)

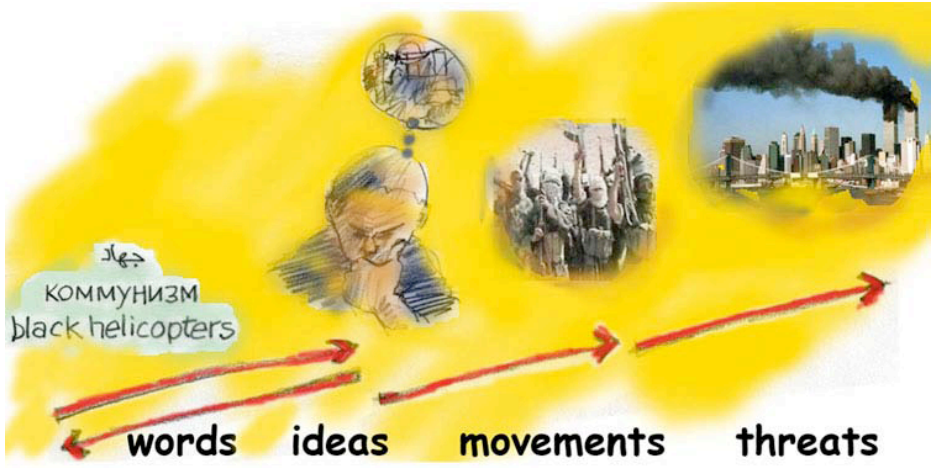
Ideology	Author	No. of text samples
Marxism-Leninism	Lenin	155
National Socialism (Nazism)	Hitler	83
Palestinian nationalism, armed overthrow of Israel	Al-Aqsa Martyrs Brigade	1
Islamism, global Salafism	Bin Laden	1
Islamism, destruction of Israel	HAMAS	3
Kahanism	Kahane Chai (Kach)	1
Mahdaviat, elimination of Israel	Ahmadinejad	2
Palestinian nationalism, violent overthrow of Israel	Palestinian Islamic Jihad	2
Irish Republicanism, armed overthrow of British rule	Real IRA	2
SUBTOTAL (hostile ideologies - 10%)		250
None	Randomly selected from WWW	2,250
TOTAL (all documents - 100%)		2,500

(Documents are in multiple languages.)

Experiments:

- Create feature vectors from all 2,500 documents using PARAFAC2 term-by-concept matrices
- Train a classifier to use vectors to distinguish between:
  - ideological and non-ideological
  - Marxism-Leninism and Nazism

Ideological Classification



Hypothesis: there could be a link between *religious* texts and threats

250 Ideological documents (Hitler, Lenin, etc.)
2250 Other web documents

→ Learn concept space with PARAFAC2, then train ensemble decision tree classifier

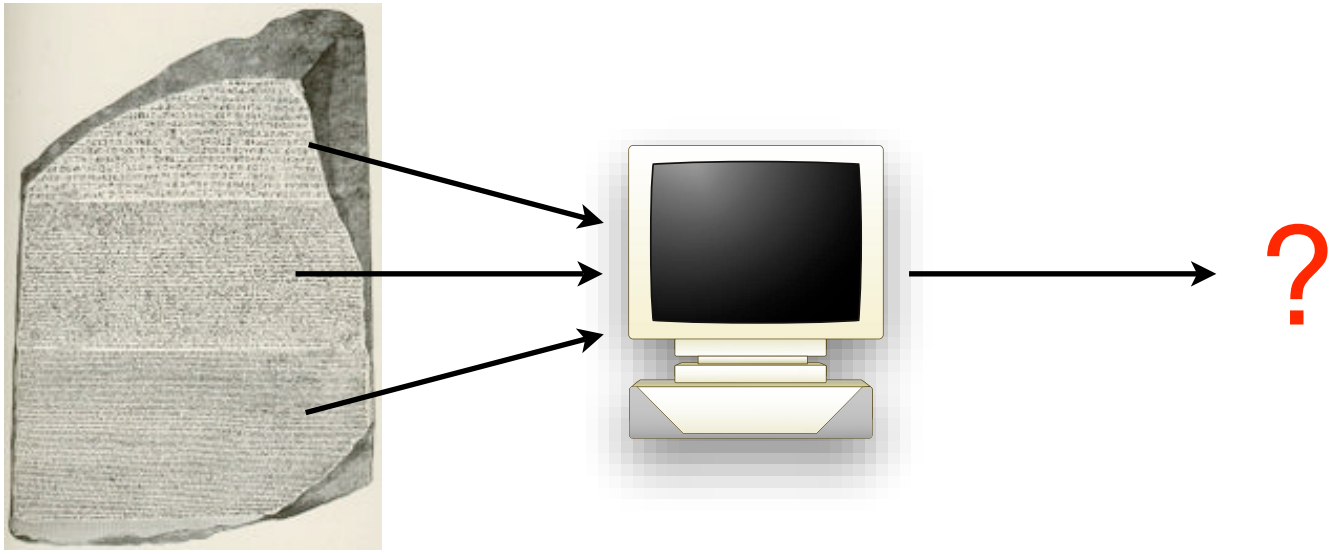
	Baseline accuracy	Actual accuracy (10-fold cross-validation)
'Hostile ideology' versus not	90.0%	98.9%
Marxism-Leninism versus Nazism	65.1%	94.7%
For comparison: movie reviews (+/-)	50.0%	64.9%

It turns out that the Bible is apparently a significantly better prism through which to look at ideologies than to look at movie reviews!

Multilingual Sentiment Analysis

(Bader, Kegelmeyer, Chew, 2011)

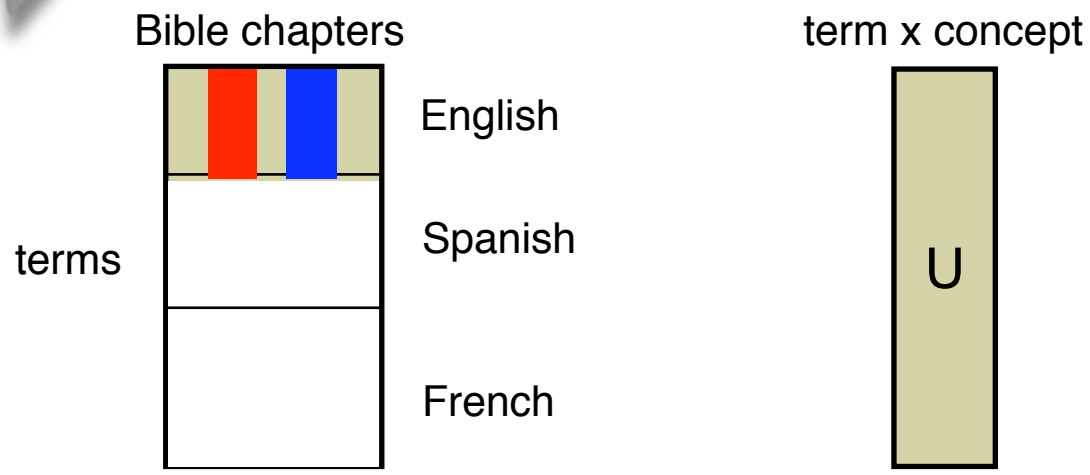
When starting solely from an English sentiment lexicon, can we classify sentiment in other languages?



- Label English chapters of Bible according to emotional valence or +/- sentiment
- Obtain language-independent features
- Train classifier
- Test on other languages

English Sentiment Classification

(Bader, Kegelmeyer, Chew, 2011)



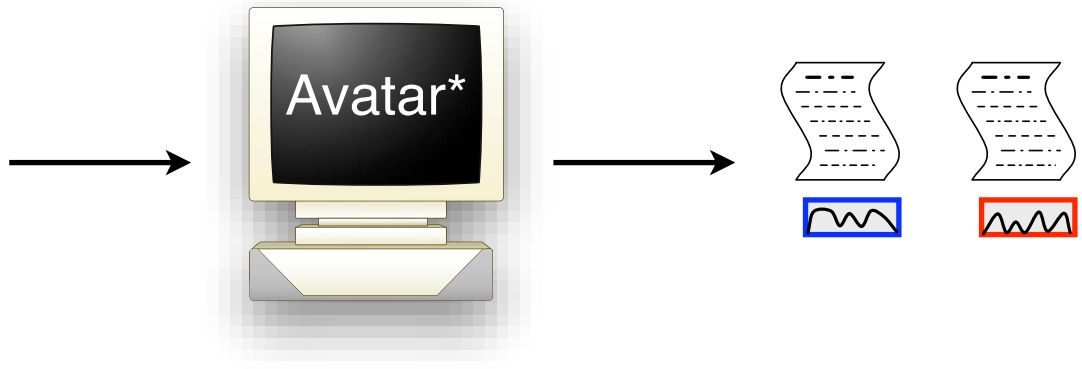
Project documents

- Project chapters into LSA concept space to get document feature vectors
- Train ensemble decision tree on feature vectors to classify sentiment
- Details on how we avoid learning topics are in our paper

Document feature vectors

dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

Train predictive model

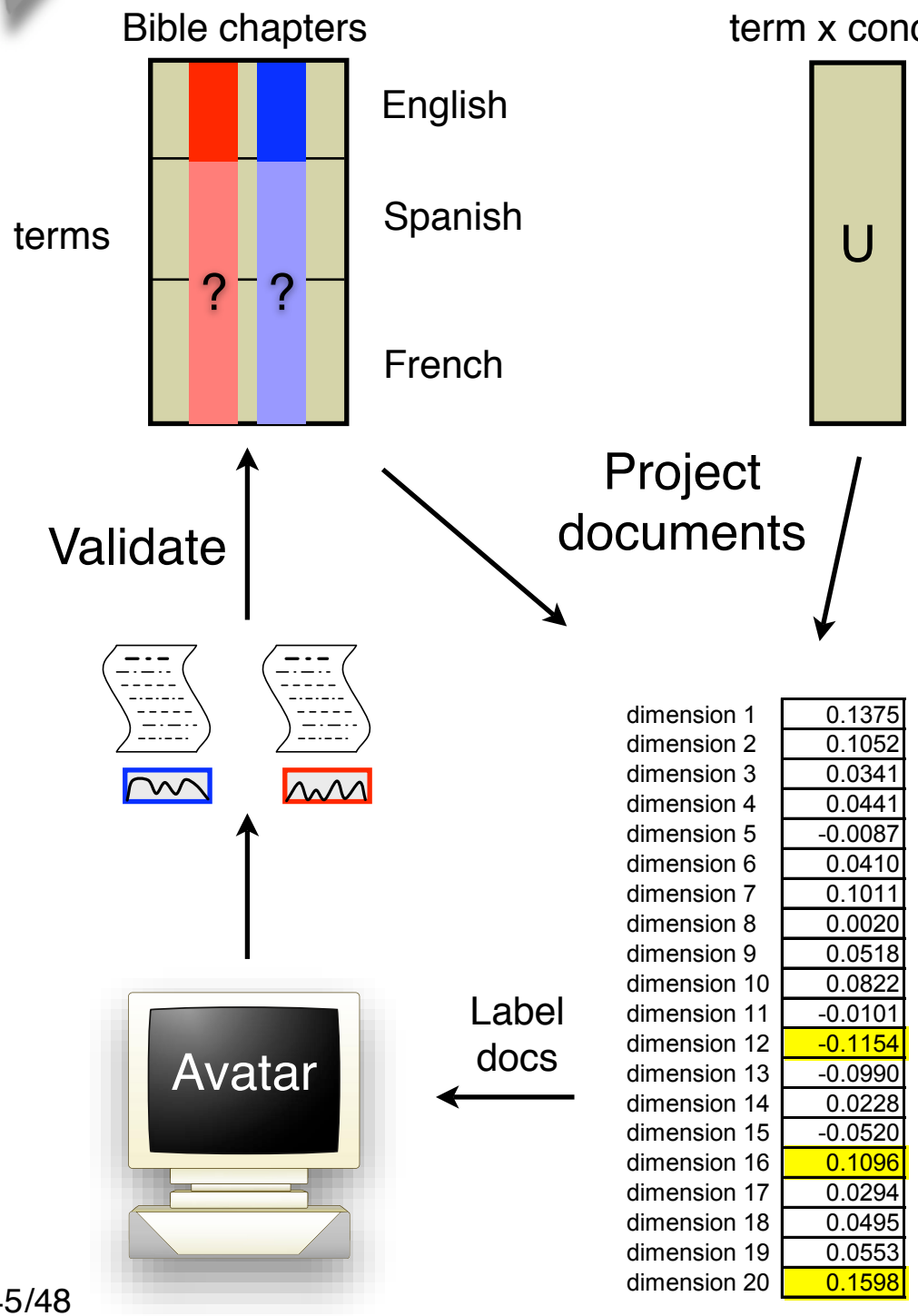


* Avatar = ensemble decision tree software (Sandia)

Sandia National Laboratories

Validation on Foreign Languages

(Bader, Kegelmeyer, Chew, 2011)



- Obtain feature vectors for the 200 chapters in other languages
- Use classifier to label chapters
- Validate with labels from English

72% accuracy in French, Spanish, German

Discussion

- We have an effective statistics-based method for comparing and making sense of documents in any of 54 languages, including all the world's major languages

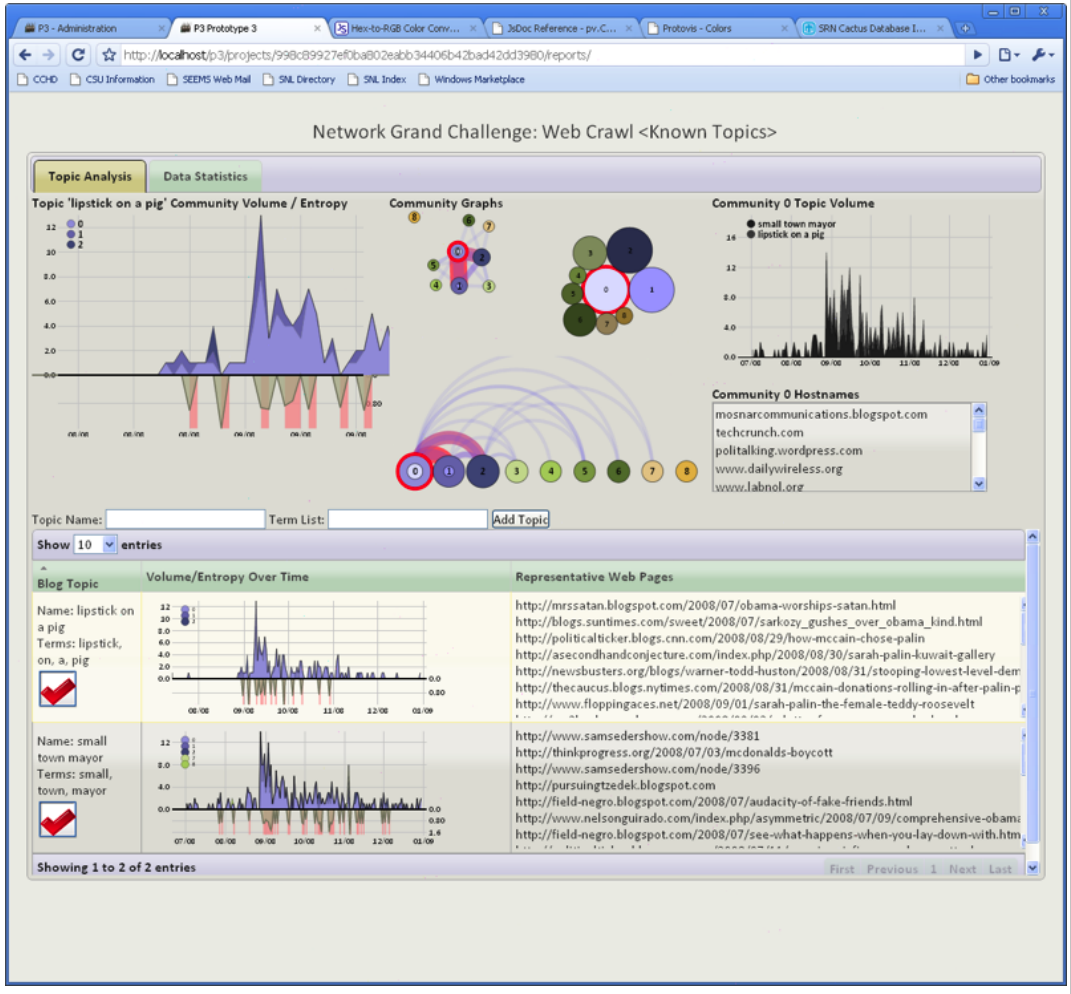
- Language morphology helps performance
  - Can deal with some out-of-vocabulary terms
  - Term alignment improves the associations made by SVD
  - LMSATA gets multilingual precision close to 90%

- Multilingual framework provides a means for various analyses
  - Document similarities and clustering
  - Ideological classification
  - Sentiment analysis

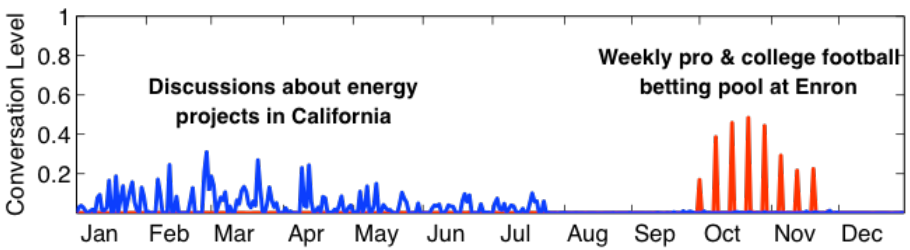
Related Text Analysis Projects

- Discussion tracking in emails
- Uncovering plots in text (scenario discovery)
- Network data exfiltration analysis
- Higher-order web link analysis
- Unsupervised part-of-speech tagging
- Identifying emerging keywords of interest

Analysis tools for web forecasting



Identifying unusual activity in Enron emails



California, power, utilities, energy, utility, governor, market

games, week, missed, picked, prize, wins, scored, upsets

Executives and project managers

2 dozen employees

Selected References

- Bader, Kegelmeyer, and Chew (2011) Multilingual sentiment analysis using latent semantic indexing and machine learning. Submitted to ACL-HLT.
- Chew et al. (2011) An information-theoretic, vector-space-model approach to cross-language information retrieval, *Natural Language Engineering*.
- Bader and Chew (2010) "Algebraic Techniques for Multilingual Document Clustering," in *Text Mining: Applications and Theory,* Wiley.
- US Patent Application No. 12/352,621 filed January 13, 2009. "Technique for Information Retrieval Using Enhanced Latent Semantic Analysis," Peter Chew and Brett Bader.
- Bader and Chew (2008) Enhancing multilingual latent semantic analysis with term alignment information. *COLING 2008.*
- Chew, Bader, and Abdelali (2008) Latent Morpho-Semantic Analysis: Multilingual Retrieval with character N-grams and mutual information. *COLING 2008*.
- Chew, Kegelmeyer, Bader and Abdelali (2008) The Knowledge of Good and Evil: Multilingual Ideology Classification with PARAFAC2 and Machine Learning. *Language Forum* (34), 37-52.
- Chew, Bader, Kolda and Abdelali (2007) Cross-language information retrieval using PARAFAC2. *Proceedings of KDD 2007.*
- Chew and Abdelali (2007) Benefits of the 'massively parallel Rosetta Stone': cross-language information retrieval with over 30 languages, *Proceedings of the Association for Computational Linguistics conference,* 2007.

Brett Bader (bwbader@sandia.gov)

http://www.sandia.gov/~bwbader